




3 1761 10374388 6





Digitized by the Internet Archive  
in 2023 with funding from  
University of Toronto

<https://archive.org/details/31761103743886>













12-001



Statistics Canada Statistique Canada

---

# **SURVEY METHODOLOGY**

---

**June 1984**

---

**Volume 10**

---

**Number 1**

---

**SPECIAL EDITION**  
**Analysis of Survey Data**  
**— Issues and Methods**



---

A Journal produced by  
Statistics Canada

---

**Canada**





Special Edition

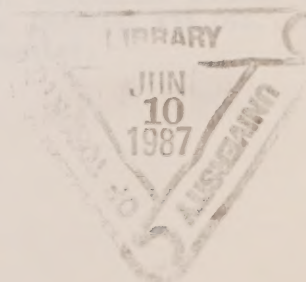
A Journal produced by Statistics Canada

C O N T E N T S

Preface.....	0
On Analytical Statistics from Complex Samples LESLIE KISH.....	1
An Introduction to Linear Models and Generalized Linear Models: Concepts and Methods DAVID A. BINDER.....	8
Adjusting Sub-Annual Series to Yearly Benchmarks PIERRE A. CHOLETTE.....	35
Examining Expenditures on Energy LOUISE A. HESLOP.....	50
Logistic Regression Analysis of Labour Force Survey Data S. KUMAR and J.N.K. RAO.....	62
Application of Linear and Log-Linear Models to Data from Complex Samples ROBERT E. FAY.....	82
Least Squares and Related Analyses for Complex Survey Designs WAYNE A. FULLER.....	97
Selected Bibliography of Data Analysis for Complex Surveys.....	119

8-3200-501  
Reference No.  
Z - 079

ISSN: 0714-0045







## SURVEY METHODOLOGY

June 1984

Vol. 10

No. 1

### Special Edition

A Journal produced by Statistics Canada

---

Editorial Board:	R. Platek	- Chairman
	M.P. Singh	- Editor
	G.J.C. Hole	
	C. Patrick	
	P.F. Timmons	
	H. Lee	- Assistant Editor

---

#### Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed; however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department.

#### Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested.





## PREFACE

This issue is devoted to presenting papers given at a symposium entitled **Analysis of Survey Data - Issues and Methods**, held at Statistics Canada on Thursday May 3, 1984.

The symposium was jointly sponsored by Methodology Research Committee at Statistics Canada and the Laboratory for Research in Probability and Statistics at Carleton and Ottawa Universities. The aim of this symposium was to demonstrate how recent developments in the area of analysis of data from complex surveys could be applied to analytic studies in Statistics Canada.

The symposium opened with remarks from the Chief Statistician, Martin B. Wilk, who emphasized the importance that Statistics Canada places in enhancing its research and development capacity and in the joint endeavours by the practitioners and academics on such issues. The symposium consisted of two sessions: - A morning session, chaired by Leslie Kish of the Institute for Social Research at the University of Michigan, which included contributions from Statistics Canada presented by D. Binder, P. Cholette, L. Heslop and S. Kumar, in addition to the presentation of an overview of the analysis issues by the Chairman.

The afternoon session chaired by the Deputy Chief Statistician, Ivan P. Fellegi started with brief remarks from the chair and included papers from R. Fay, U.S. Bureau of the Census and W. Fuller, Iowa State University. The session concluded with general discussion of the developments on the data analysis issues led by J.N.K. Rao, Carleton University. Well over 200 participants from various Universities and Federal and Provincial Government Departments attended the symposium.

A selected bibliography on the topic compiled by the Project Team on the Analysis of Data from Complex Surveys is also given at the end.



ON ANALYTICAL STATISTICS FROM COMPLEX SAMPLES<sup>1</sup>Leslie Kish<sup>2</sup>

I want to plead the case that an important and urgent task facing mathematical statistics consists of providing useful expressions for analytical statistics for complex sample designs. I should like to describe these problems to mathematical statisticians who should find them interesting because they meet the criteria of all good problems: they are important, unsolved and solvable.

The most important and difficult problems of survey sampling still await adequate mathematical treatment: the textbooks are aimed almost entirely at producing good estimates of aggregates, means and ratio means. One may also deal with the differences of two of these, but there is only fleeting and occasional reference to this problem. However, with that we come to the end of the statistical tools available for complex samples.

As sampling theory developed, probability sampling has been capturing the field of respectable sampling practice with sample designs, which are often simultaneously economical and complex. One result has been an increasing volume of sample survey data which is of high quality and which researchers wish to put to more involved analytical use. But the mathematical statistics for doing this validly are lacking. The available analytical statistics assume independence among the selected elements: but this independence is lacking in complex sample designs. Thus the researcher may be forced to forego the analysis which he considers desirable and valuable. But if he is too impatient or too ignorant for that act of self-denial, he may go ahead and use the srs formulas he finds in books on statistics, which often result in very serious errors.

I hope that mathematical statisticians will be impressed with the importance of the unsolved problems of analytical statistics for data arising from complex sample designs. The lack of these is a more frequent source of gross mistakes than any other kind of departure from the usual assumptions.

---

<sup>1</sup> Overview talk for the symposium.

<sup>2</sup> Leslie Kish, Institute for Social Research, The University of Michigan.



These problems are important, unsolved and interesting. You may ask: are they solvable now? Supporting my affirmative answer are three sources of justification. First, we observe the great recent advances in statistical theory. Secondly, the rapid increases in the quantity and quality of electronic computing machines make the time ripe for the solution of some of these problems. There is new interest in a general method which holds promise of rapid advance toward useful approximations. At the Survey Research Center we are now introducing this method for computing estimates of variances for regression coefficients and other statistics for which formulas are not now available.

It seems to me that this procedure resembles that of Alexander when he "solved" the Gordian knot. From a theoretical viewpoint I don't know whether it constitutes a solution of the problem or its avoidance. But insofar as it promises to give good approximations for much needed variances the practicing statistician will welcome its development with enthusiasm and interest. In this way one may obtain estimates of the confidence intervals of some analytical statistics for which specific formulas are not now available.

All of the above is verbatim from my talk to a joint session of the American Statistical Association and the Institute of Mathematical Statistics in 1957. Since then our situation has changed but little. Our 1957 hopes for that cut of the Gordian knot is now much used as BRR or Balanced Repeated Replications (Kish and Frankel 1970, 1974). But my moving plea for distribution theory for doubly complex analytical statistics did not move the mathematical statisticians. I know now why not, since I am sadder and wiser now. First, statisticians like other scientists work not on what solutions are needed but on those that seem feasible at the time. (Like nuclear bombs, for example.) Second, distribution theory for complicated statistics for complex samples seems too difficult to solve. Third, the solutions would have too many parameters to be useful. Thus my views in (Kish 1978) and today are more sober: "New computational methods can give us approximate variances that appear satisfactory for practical purposes. However, it would be more satisfying to have mathematical distribution theory for analytical statistics (e.g. regression coefficients) without the assumptions of independence, but with complex correlations between sample observations. We may hope for some progress, but not for generally useful results, because of mathematical

complexities, and even more because the numbers of needed parameters will prove too great for practical utility."

Here follow seven important points about complex samples put boldly. They are not all widely known or believed, but I ask you to know, believe, use and teach them, as I do.

1. The effects of complex designs must be considered separately for point estimates and for probability statements, like confidence intervals or tests of hypotheses. For point estimates we have for all sample designs consistent approaches to parameters from similar probability-weighted (H-T) estimators. But the probability statements like confidence intervals are highly subject to design effects, especially in cluster sampling.

a) "Statistics (means, regression coefficient, etc.) approach their population values as the sample size increases.

b) The approach is generally slowed by design effects.

c) The design effects differ for different statistics, for different variables and different sample designs." (Kish and Frankel, 1974).

That paper also presents the most convincing evidence for these points; and evidence is widespread; e.g. (Verma et al 1980). Nevertheless two famous statisticians completely misstated our position in discussions of our paper: "Here the authors make the important observation that the confidence interval statements for the unknown parameter are numerically not much affected by the lack of independence of observations introduced by complex survey techniques such as stratified cluster sampling." Alas, that mistake gets quoted by other theoreticians who fail to read our answer of survey samplers: "They misunderstand completely our principal and repeated message: that confidence interval statements are numerically greatly affected by the lack of independence of observations introduced by complex survey techniques such as stratified cluster sampling." (Kish and Frankel, 1974).

This misunderstanding shared by naive non-statisticians with sampling theorists causes troubles for us survey samplers: hence we are working on a clearer statement.

2. Do we need sampling errors for analytical statistics for data from complex surveys? Or have a few of us been devoted to a negligible even trivial problem? I feel like a St. Sebastian, the target practice for the

slings and arrows of diverse outrageous heathen. (Mixed metaphors are better than fixed or random.) First come the market researchers and pollsters who ignore us, though some have learned to put a  $\sqrt{pq/n}$  between a 2 and (pq/n). Second, some demographers write that with their large samples and larger measurement errors they have no time for sampling errors. Third come the mathematical psychologists, econometricians and biometricians who take their linear models straight from mathematical statistics, and that hurts. Fourth, even more hurtful are the mathematical statisticians themselves, who either forget that their n's do not justify their means, or they invoke IID, or they use some Bayesian exorcism against the spirits of the sample design. Fifth and worst are sampling theorists who display theorems to prove that, with completely specified models of arbitrary superpopulations, we need not worry about whence or how our elements were selected, nor weight them for unequal selection probabilities. They even convince a few survey samplers that they can dwell on some Olympus with their models and not come down to earth where the population lives.

From these necessarily brief remarks you notice that I am an extremist for several reasons: a) Design effects for analytical statistics provide common evidence for imperfectly specified models for the best stratified samples; b) We frequently find the effects of selection weights on samples; c) Relations between predictor and predictand variables exist in actual individuals, and they in real populations, and these interact with sample designs. (I am developing these points in a book on Statistical Design for Social Research for Wiley, 1985.)

My philosophy is consistent, but in practice I am less dogmatic. I recognize that in practice: a) it is never possible to cover completely our target populations, hence we must always resort to models for inference; b) probability sampling is too costly and not feasible for most experiments; c) despite lack of randomization either in selection or in treatments, we often blunder our way to reliable results with care, replication, design, additivity and a little bit of luck.

3. Analytical statistics begin with subclasses and with their comparisons. In the last three decades much useful material has been published about variances and design effects for subclasses. There are masses of empirical results and several useful guiding rules based on them (Kish 1980, Kish et

al. 1976, Verma et al. 1980), also some recent theory (Rust 1984, Chapter 6).

a) Distinguish between proper domains and the more common crossclasses, on which we focus here.

b) Selection probabilities are preserved for crossclasses but sample sizes become highly variable.

c) Estimates of totals and means from complex samples are retained in ratio and conditional forms.

d) Design effects for crossclasses tend to approach to almost 1 proportionately as the subclass sizes per primary cluster approach 1. This approximate model needs care and qualification but it is preferable to all venerable alternatives about design effects: that it is simply 1, or some other constant, or the same as for the entire sample. The pooled model may be often better than separate and highly variable computations.

4. Comparisons of paired means tend to have design effects greater than 1 but considerably less than the sum of the two variances. These reductions due to positive covariances (hence to a kind of additivity) have been found widely and regularly for comparisons both of crossclasses and of periodic surveys (Kish 1965, 14.1, also the above).

5. For complex analytical statistics several methods exploit the potentialities of electronic computing: Taylor linearized (delta) methods, including machine differentiation, Balanced Repeated Replications and Jackknife Repeated Replications, all have been shown to yield useful estimates of variance and design effects for complex samples (Kish and Frankel 1970 and 1974; Woodruff and Causey 1978), Bootstrapping may also be added in the future (Rao 1984).

Analytical statistics consistently show design effects greater than 1, significantly greater in every sense, but also lower than design effects for means. The relations of design effects between diverse coefficients and comparisons with those for means show some regularities.

For useful guidance we need not only more empirical work but also more results from sampling theory and model building. I am disappointed frankly that since our early work we have not seen more publications in theory and models that would be directly useful for guiding inference for actual data. The empirical bases of design effects are necessary, but to satisfy our intellectual needs for understanding we need more theory and better models.



Furthermore, even our practical needs remain unsatisfied with merely empirical design effects, because they are functions jointly of the variables, of the type of estimates, of the sample design used and of the population basis for the data. That four-dimensional source of variation is too complex and we need theory to construct models for greater simplicity.

6. Categorical data analysis is an important area, rapidly developing, and several contributions have been made to apply these methods to complex survey data (Fay 1982; Landis et al. 1982; Koch et al. 1975). These also have implications for analysis of variance where some of the earliest models were started, but not followed (Kempthorne and Wilk, 1955; Tukey and Cornfield).

7. As for the future I am hopeful about contributions from theory to applications but for two exceptions. First, mathematical statistics has not and will not give us complete distribution theories that will be useful directly, because there are too many parameters in the double complexity of analytical statistics from complex surveys. Second, model builders cannot make those complexities vanish. They will however guide us toward better and more comprehensive inference. Also toward better utilization and presentation of analytical statistics from complex surveys.

## REFERENCES

- [1] Fay, R. (1982). Contingency table analysis for complex sample designs: CPLX. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 44-53.
- [2] Kish, L. (1957). Some unsolved problems of complex samples. Paper for Joint Meeting of the American Statistical Association and Institute for Mathematical Statistics.
- [3] Kish, L., and Frankel, M.R. (1970). Balanced repeated replications for standard errors. JASA, 65, pp. 1071-94.
- [4] Kish, L., and Frankel, M.R. (1974). Inference from complex samples. JRSS (B), 36, pp. 1-74.

- [5] Kish, L. (1980). Design and estimation for domains. The Statistician (London), 29, pp. 209-22.
- [6] Kish, L., Groves R.M., and Krotki (1976). Sampling errors for fertility surveys. Occasional paper 17, London: World Fertility Surveys, 61 pages.
- [7] Koch, G., Freeman, D., and Freeman, J. (1975). Strategies in the multivariate analysis of data from complex surveys. International Statistical Review, 43, pp. 53-59.
- [8] Landis, J.R., Lepkowski, J., Eklund, S., and Stehouwer, S. (1982). A statistical methodology for analyzing data from a complex sample survey. Vital and Health Statistics, Series 2 - No. 92. DHHS Publ. No. 82-1366. Public Health Service, Washington, U.S. Government Printing Office.
- [9] Rao, J.N.K. (1984). Bootstrap inference with stratified samples. (Submitted for Publication).
- [10] Rust, K.F. (1984). Techniques for Estimating Variances for Sample Surveys. The University of Michigan, Ph.D. dissertation.
- [11] Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. JRSS (a) 143, pp. 431-73.

## AN INTRODUCTION TO LINEAR MODELS AND GENERALIZED LINEAR MODELS: CONCEPTS AND METHODS

David A. Binder<sup>1</sup>

Univariate statistical models, linear regression models and generalized linear models are briefly reviewed. Examples of a two-way analysis of variance, a three-way analysis of variance and logistic regression for a three way layout are given.

### 1. INTRODUCTION

The purpose of this presentation is to give a bird's-eye view of some of the concepts used in statistical applications for modelling data

The use of data sampled from a population to estimate means and proportions is now a common practice. In Section 2 we briefly review this concept and describe the interval estimates obtained from constructing confidence intervals.

Linear regression and analysis of variance models are often used to reduce multi-dimensional data to a model consisting of a few parameters. This tool is a valuable device for the analyst looking for a deeper understanding of a complex data set. These methods are reviewed in Section 3.

The concepts of linear regression methods can be extended to a much wider class of models through the generalized linear models described by Nelder and Wedderburn (1972). This is particularly useful when the dependent variable is categorical as opposed to continuous. In Section 4 we review the structure of these models.

Brief mention of appropriate diagnostics to guard against model failure and to detect multicollinearities is given in Section 5.

---

<sup>1</sup> David A. Binder, Institutional and Agriculture Survey Methods Division, Statistics Canada.

## 2. UNIVARIATE MODELS

### 2.1 Binomial Models

Suppose we have a large population from which we will select a sample and we take an observation from each selected unit. If the sample size is  $n$ , we denote the observations by  $Y_1, Y_2, \dots, Y_n$ . The purpose of collecting this data is that we would like to make some inferences about the population based on this sample. For example, our population could be residents of Canada and our data are defined as

$$Y_j = \begin{cases} 1 & \text{if the person was born in Canada} \\ 0 & \text{if the person was born outside of Canada,} \end{cases}$$

for the  $j$ -th individual selected. Based on this sample we would like to make some inferences on the proportion of people in the population who were born in Canada.

If a simple random sample of  $n = 5000$  residents is selected and the actual proportion of persons born in Canada is  $p = 0.85$ , then the number of persons in our sample who are born in Canada will be a random variable with a binomial distribution given by

$$f(y) = \binom{5000}{y} (.85)^y (.15)^{5000 - y}; y = 0, 1, \dots, 5000.$$

In this case, since we know  $p = .85$ , we can completely describe the properties of  $Y = \sum Y_j$ , the total in our sample who are born in Canada. For most statistical applications, though, we do not know all the characteristics of the population and we use our sample to make inferences about this population. For example, suppose we do not know the value of  $p$  in the previous example. Then we can say that the number of persons in our sample who were born in Canada will be a binomial random variable having a distribution given by

$$f(y) = \binom{5000}{y} p^y (1 - p)^{5000 - y}; y = 0, 1, \dots, 5000.$$



Now, the usual estimator for  $p$ , based on this data is  $\hat{p} = \bar{Y} = \sum Y_{.j}/5000$ . We let  $s(\hat{p}) = \{\hat{p}(1-\hat{p})/(5000)\}^{\frac{1}{2}}$ . This is our estimate of the standard error of  $\hat{p}$ . Now, it turns out that  $\hat{p} \pm 1.96 s(\hat{p})$  is a random interval which has a 95% chance of including the true unknown value of  $p$ . This interval is called a 95% confidence interval. By changing the value of 1.96 we would either shorten or lengthen the confidence interval, thus changing the coefficient from 95% to some other value. These coefficients can be obtained from probabilities associated with the standard normal distribution.

We have described the binomial model via a simple random sample from a large population. Thus, all our inferences pertain to that population. However, in many contexts we would like our inferences to relate to other populations which we believe have been generated under similar conditions. For example, the number of deaths in Canada from a particular age-sex group in a given year may be thought of as a single realization from a binomial model, where each individual has the same probability of dying and the individual deaths are essentially independent. If this probability of dying is constant over a number of years then the number of deaths in one year can be used to make inferences for other years, even though the populations are different. (Life insurance companies and their actuaries rely on these types of assumptions in their calculations.) Providing that individual deaths are independent, assumptions about constancy of the probability of death are testable using these binomial models.

It should be pointed out that by using some generalized linear models to be described in Section 4, it may be possible to improve on the assumption of constant probabilities for all individuals, by allowing the probabilities to depend on other factors such as age, sex, health status, smoking habits, weight, etc.

## 2.2 Normal Models

An important distribution used in modeling data is the normal distribution given by

$$f(y) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (y - \mu)^2 \right\}; \quad -\infty < y < \infty.$$

The population mean is  $\mu$  and is usually the parameter of interest. The population variance is  $\sigma^2$ .

If we observe data  $Y_1, Y_2, \dots, Y_n$  from this population, our usual estimator for  $\mu$  is  $\hat{\mu} = \bar{Y} = \sum Y_j / n$ . Our estimator for the standard error of  $\hat{\mu}$  is given by  $s(\hat{\mu}) = s/n^{1/2}$ , where

$$s^2 = \sum (Y_j - \hat{\mu})^2 / (n - 1).$$

As in the case of the binomial model, for large samples the 95% confidence interval is given by  $\hat{\mu} \pm 1.96s(\hat{\mu})$ . This is a random interval which has a 95% chance of including the true value of  $\mu$ . For small samples (e.g.  $n < 60$ ), the value 1.96 may be replaced by the appropriate value from the t distribution for more accurate intervals. Other confidence coefficients may also be obtained by changing the value 1.96 to the appropriate percentile from the standard normal or t distribution.

In some applications, the assumption of constant variance is unrealistic, particularly in the linear models to be discussed in Section 3. A simple extension of this model is to assume that the variance of  $X_i$  is given by  $\sigma_i^2$  where  $\sigma_i^2 = \sigma^2/w_i$ . Here we assume that  $w_1, w_2, \dots, w_n$  are known weights. In this case  $\hat{\mu} = \sum w_j Y_j / \sum w_j$ , a weighted average of the data. Also  $s(\hat{\mu}) = s/(\sum w_j)^{1/2}$ , where

$$s^2 = \sum w_j (Y_j - \hat{\mu})^2 / (n - 1).$$

Confidence intervals for  $\mu$  are obtained analogously. It should be pointed out here that the weights,  $w_1, \dots, w_n$  are based on the normal model specification and are usually unrelated to sampling weights which are derived from complex survey designs from finite populations. When fitting models to finite populations based on data from a complex survey design, the analyst may wish to incorporate both the model weights as well as the sampling weights in the estimation.

### 2.3 Exponential Family Models

The binomial and normal models just described can be viewed as special cases of a much wider class of models known as the exponential family. The general form which we will use for this model is given by:

$$f(y_j) = \exp[\kappa_j \{y_j \theta - b(\theta)\} + c(y_j, \kappa_j)],$$

where  $y_j$  takes values which do not depend on  $\theta$ .

We assume  $\kappa_j = \kappa w_j$  where  $w_1, \dots, w_n$  are known. In many cases  $\kappa$  will also be known.

#### Example 1 (Binomial Proportion)

We let  $\bar{y}_j = y_j/n_j$  be the sample proportion from a binomial model based on  $n_j$  observations. Therefore we have:

$$f(\bar{y}_j) = \binom{n_j}{n_j \bar{y}_j} p^{n_j \bar{y}_j} (1-p)^{n_j(1-\bar{y}_j)}; \bar{y}_j = 0, \frac{1}{n_j}, \frac{2}{n_j}, \dots, 1,$$

$$E(\bar{y}_j) = p, \text{Var}(\bar{y}_j) = p(1-p)/n_j,$$

$$\theta = \log[p/(1-p)].$$

$$\kappa_j = n_j.$$

$$b(\theta) = \log(1 + e^\theta).$$

#### Example 2 (Normal)

Suppose  $y_j$  is normally distributed with mean  $\mu$  and variance  $\sigma_j^2$ . We have:

$$f(y_j) = (2\pi\sigma_j^2)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \left( \frac{y_j - \mu}{\sigma_j} \right)^2 \right\}; -\infty < y_j < \infty$$

$$E(y_j) = \mu, \quad \text{Var}(y_j) = \sigma_j^2,$$

$$\theta = \mu,$$

$$\kappa_j = 1/\sigma_j^2,$$

$$b(\theta) = \mu^2/2.$$

### Example 3 (Poisson Mean)

Suppose  $y_j$  is Poisson with mean  $n_j\lambda$ . Letting  $\bar{y}_j = y_j/n_j$ , we have:

$$f(\bar{y}_j) = e^{-n_j\lambda} (n_j\lambda)^{n_j\bar{y}_j} / (n_j\bar{y}_j)!; \quad \bar{y}_j = 0, \frac{1}{n_j}, \frac{2}{n_j}, \dots,$$

$$E(\bar{y}_j) = \lambda, \quad \text{Var}(\bar{y}_j) = \lambda/n_j,$$

$$\theta = \log \lambda,$$

$$\kappa_j = n_j,$$

$$b(\theta) = e^\theta.$$

### Example 4 ( $\chi^2$ )

Suppose  $y_j$  has a  $\sigma^2\chi_{v_j}^2/v_j$  distribution. This is common for analysis of variance and variance components models, where  $y_j$  is the mean-square. Then, we have:

$$f(y_j) = y_j^{(v_j-2)/2} \left(\frac{v_j}{2\sigma^2}\right)^{v_j/2} \exp\{-y_j v_j/(2\sigma^2)\} / \Gamma(v_j/2); \quad y_j \geq 0,$$

$$E(y_j) = \sigma^2, \quad \text{Var}(y_j) = 2\sigma^4/v_j,$$

$$\theta = -1/\sigma^2,$$

$$\kappa_j = v_j/2,$$

$$b(\theta) = -\log(-\theta).$$

As we can see from these examples, the exponential family includes a wide variety of common distributions. In general, we have



$$E(y_j) = b'(\theta) = \mu, \quad \text{Var}(y_j) = b''(\theta)/\kappa_j = V_j$$

where  $b'(\cdot)$  and  $b''(\cdot)$  denote the first and second derivatives of  $b(\cdot)$ .

If  $y_1, \dots, y_n$  are independent, then the maximum likelihood estimate of  $\theta$  is given by the solution to:

$$\hat{\mu} = \sum \kappa_j y_j / \sum \kappa_j = \sum w_j y_j / \sum w_j$$

where  $\hat{\mu} = b'(\hat{\theta})$ . This implies that there is a large family of models where a weighted sample mean provides an efficient estimator of the population mean. The estimated variance of  $\hat{\mu}$  is given by

$$\begin{aligned} \hat{V}(\hat{\mu}) &= (\sum \kappa_j^2 \hat{V}_j) / (\sum \kappa_j)^2 \\ &= b''(\hat{\theta}) / (\sum \kappa_j). \end{aligned}$$

For large samples, the 95% confidence interval for  $\mu$  is given by  $\mu \pm 1.96 \times \{\hat{V}(\hat{\mu})\}^{1/2}$ , providing the model is true.

In cases where  $\kappa_j = \kappa w_j$  is known only up to the constant of proportionality  $\kappa$ , (e.g. normal model), it will be necessary to estimate the value of  $\kappa$ . The maximum likelihood estimate is given by the solution to:

$$\sum w_j [y_j \theta - b(\theta) + \frac{\partial c(y_j, \kappa_j)}{\partial \kappa_j}] = n.$$

Alternatively, an unbiased estimator for  $\hat{V}(\hat{\mu})$  which is less model-dependent is given by

$$\hat{V}_1(\hat{\mu}) = \frac{\sum w_j (y_j - \hat{\mu})^2}{(n-1)(\sum w_j)}.$$

This may be used instead to create the confidence intervals for  $\hat{\mu}$ .

The

main assumption required for the validity of this approach is that  $\text{Var}(y_j) \propto 1/w_j$ .

### 3. LINEAR MODELS

#### 3.1 One Way Analysis of Variance

A simple extension of the univariate normal models, described in Section 2.2, is the one-way analysis of variance (ANOVA) model. Here, in addition to observing one characteristic from each individual sampled, we also have a sub-population identifier. Some such identifiers could be age-sex groups, industry/occupation groups, etc. Here the model could be written as

$$y_{ij} = \mu_i + \varepsilon_{ij}; i = 1, \dots, I; j = 1, \dots, n_i,$$

where the  $\mu$ 's are population means, which differ among subpopulations and the  $\varepsilon$ 's are assumed to be independent normal with variances  $\sigma_{ij}^2 = \sigma^2/w_{ij}$ , where the  $w_{ij}$ 's are known weights. In most applications the weights are constant.

The usual estimator for  $\mu_i$  in this model is

$$\hat{\mu}_i = \sum_j w_{ij} y_{ij} / \sum_j w_{ij}.$$

Under the model assumptions, the estimated means are independent normal with  $E(\hat{\mu}_i) = \mu_i$  and  $\text{Var}(\hat{\mu}_i) = \sigma^2 / \sum_j w_{ij}$ . From this, confidence intervals for the individual means may be derived.

An alternative but equivalent description of this model is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where  $\sum \sum w_{ij} \alpha_i = 0$ . Here we have

$$\mu = \sum \sum w_{ij} \mu_i / \sum \sum w_{ij}$$

$$\alpha_i = \mu_i - \mu.$$

An extension of this representation is particularly useful for two-way and higher order analysis of variance models, to be discussed in Sections 3.2 and 3.3. One of the main questions of interest for these models is whether all the means are equal. This is equivalent to  $\mu_1 = \mu_2 = \dots = \mu_I$  or  $\alpha_1 = \alpha_2 = \dots = \alpha_I = 0$ . Standard ANOVA statistical packages (e.g. SAS, SPSS, etc.) are available to test these hypotheses. A related problem is: Which subpopulation means are equal, given that we have concluded already that not all means are equal? When we have no further structure (such as in a two-way ANOVA), this is known as the multiple comparison problems. Special treatments for this problem are available in many statistical packages.

### 3.2 Two-Way Analysis of Variance

The data of Table 1 has been taken from the 1975 Sri Lanka Fertility Survey (see Little, 1982). The cell means describe the average number of children ever born cross-classified by Marital Duration and Level of Education.

The row and column means seem to indicate that the average number of children increases with longer marriage durations and decreases with more schooling. Now, the two-way analysis of variance model may be written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

where the  $\epsilon$ 's are assumed to be independent normal with variances  $\sigma_{ijk}^2 = \sigma^2/w_{ijk}$ . The  $w$ 's are known weights. In most applications the weights are constant. In order to estimate the parameters of this model, it is necessary to impose constraints on these parameters, otherwise they are not unique. The usual side conditions are:

$$\sum_i \sum_j \sum_k w_{ijk} \alpha_i = 0,$$

$$\sum_i \sum_j \sum_k w_{ijk} \beta_j = 0,$$

$$\sum_i \sum_k w_{ijk} \gamma_{ij} = 0,$$

$$\sum_j \sum_k w_{ijk} \gamma_{ij} = 0.$$

The estimators are defined by the equations:

$$\sum_i \sum_j \sum_k w_{ijk} (y_{ijk} - \hat{\mu}_{ij}) \frac{\partial \hat{\mu}_{ij}}{\partial \hat{\theta}_\ell} = 0$$

where  $\hat{\theta}_1, \hat{\theta}_2, \dots$  correspondent to the parameter estimates  $\hat{\mu}, \hat{\alpha}_i$ , etc. The  $\alpha$ 's are  $\beta$ 's are referred to as main effects and the  $\gamma$ 's are the two-way interactions. This results in the following estimators:

$$\hat{\mu} = \bar{y}_{...},$$

$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...} - \frac{\sum_j \sum_k w_{ijk} \hat{\beta}_j}{\sum_j \sum_k w_{ijk}},$$

$$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...} - \frac{\sum_i \sum_k w_{ijk} \hat{\alpha}_i}{\sum_i \sum_k w_{ijk}},$$

$$\gamma_{ij} = \bar{y}_{ij.} - \bar{y}_{...} - \hat{\alpha}_i - \hat{\beta}_j,$$

where  $\bar{y}_{ij.}, \bar{y}_{i..}$ , etc. are the appropriate weighted averages.

Now, the additive model specifies that  $\mu_{ij} = \mu + \alpha_i + \beta_j$ . We have plotted the cell means from Table 1 in Figure 1. The additive model would specify that all the lines are parallel. If the data of Table 1 are fitted to the additive model, we obtain the adjusted mean values in Table 2. These are plotted in Figure 2. As we can see, the effect of the level of education has been dramatically reduced after fitting this model. This is because the more educated women were not married for as long, so that the years since first marriage proves to be the important factor. However, as the analysis of variance in Table 3 shows, all the main effects and the interactions are significant. Hence the additive model is rejected. However, only 0.4% of the total variation is explained by the Education-Marital Durations interactions, whereas 49.7% of the variation is explained by the additive model. We may surmise from this that the additive model has led to a better understanding of the data and that the Education effect is not as dramatic as it first



seemed.

### 3.3 Regression Formulation

The above analysis of variance models can be considered as special cases of the multiple linear regression model, given by

$$y_j = \beta_0 X_{0j} + \beta_1 X_{1j} + \dots + \beta_r X_{rj} + \epsilon_j,$$

where  $X_{0j}$ ,  $X_{1j}$ , ...,  $X_{rj}$  are known constants and  $\beta_0$ ,  $\beta_1$ , ...,  $\beta_r$  are unknown coefficients. We assume that the  $\epsilon$ 's are independent normal with variances  $\sigma_j^2 = \sigma^2/w_j$ , where the  $w_j$ 's are known weights. For example, in the one way analysis of variance, we could let

$$X_{0j} = 1 \text{ for all } j$$

$$X_{ij} = 1 \text{ if the } j\text{-th individual is in the } i\text{-th sub-population}$$

$$= -a_i/a_I \text{ if the } j\text{-th individual is in the } I\text{-th sub-population}$$

$$= 0 \text{ otherwise,}$$

for  $i = 1, \dots, I - 1$ , where  $a_i$  is the sum of the weights for individuals in the  $i$ -th sub-population. In this case we have

$$\mu_i = \beta_0 + \beta_i \quad \text{for } i = 1, \dots, I - 1,$$

$$\mu_I = \beta_0 - (a_1\beta_1 + \dots + a_{I-1}\beta_{I-1})/a_I.$$

Therefore  $\mu = \beta_0$  and  $\alpha_i = \beta_i$  for  $i = 1, \dots, I - 1$ .

A similar regression formulation is possible for two-way and higher order layouts as well.

Now, for the general regression model, the estimator for  $\beta_0, \dots, \beta_r$  is given by  $\hat{\beta}_0, \dots, \hat{\beta}_r$ , the solution to

$$\sum w_j (y_j - \hat{y}_j) X_{ij}, \quad i = 0, 1, \dots, r$$

where  $\hat{y}_j = \hat{\beta}_0 X_{0j} + \hat{\beta}_1 X_{1j} + \dots + \hat{\beta}_r X_{rj}$ .

In order to test hypotheses, perform model-building and develop confidence intervals for the  $\beta$ 's, we need the covariance matrix of the  $\hat{\beta}$ 's. This is given by

$$\text{Var}(\hat{\beta}) = \sigma^2 A^{-1}$$

where  $A$  is the matrix with  $(k, l)$ -th entry being  $\sum_j w_j X_{kj} X_{lj}$ . To estimate  $\sigma^2$ , we use  $\hat{\sigma}^2 = \sum_j w_j (y_j - \hat{y}_j)^2 / (n - r - 1)$ .

Many statistical packages routinely perform various hypothesis tests on  $\hat{\beta}$  using the estimated covariance matrix  $\hat{\sigma}^2 A^{-1}$  and the critical values from the appropriate F-distribution (e.g. PROC REG, PROC ANOVA and PROC GLM in SAS).

For example, Koch, Gillings and Stokes (1980) give the data in Table 4 for the number of physician visits per person per year in 1973 in the U.S. cross-classified by size of city (SMSA = Standard Metropolitan Statistical Area vs. Non-SMSA), Income (3 groups) and Education (3 groups). This data is based on the 1973 Health Interview Survey, a survey using a complex probability sample. The data are illustrated in Figure 3.

By using a regression model and performing a number of statistical tests, the following reduced model was obtained:

$$E(Y_j) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j},$$

where  $X_{1j} = 1$  if the  $j$ -th person is in an SMSA

$= 0$  otherwise,

$X_{2j} = 1$  if the  $j$ -th person has less than \$5000 family income or more than 12 years education for the family head

$= 0$  otherwise.

The estimated parameters were  $\hat{\beta}_0 = 4.18$  (standard error of 0.11),  $\hat{\beta}_1 = 0.65$  (standard error of 0.11) and  $\hat{\beta}_2 = 1.12$  (standard error of 0.09). The standard errors derived here were not those described above since the authors used the  $18 \times 18$  estimated covariance matrix from the survey to obtain the standard errors. This approach removes the assumption of independent error terms in

the model-fitting and is a common approach for analysing data from complex surveys.

In Table 5 we summarize the results. These are illustrated in Figure 4. We see that the model fit is quite good. We have reduced the data from 18 values to 3 summary statistics and also have smaller standard errors (hence higher precision) of the estimated values.

## 4. GENERALIZED LINEAR MODELS

### 4.1 Regression with a Dichotomous Dependent Variable

One of the difficulties often encountered with the linear models discussed in Section 3 is that the error terms were assumed to be normally distributed. It is true that analyses similar to those in Section 3 may be performed with non-normal errors, providing the variances of the errors still satisfy  $\sigma_j^2 = \sigma^2/w_j$  and the errors are uncorrelated. In this case the estimators we have described yield the minimum variance linear unbiased estimates of the model parameters, however better estimators (i.e. non-linear estimators) may be available. These considerations have led to generalized linear models (see Nelder and Wedderburn, 1972) and robust estimators (see Huber, 1973). We concentrate here on the generalized linear models.

For example, suppose the dependent variable,  $y_j$ , can take on only two values, 0 or 1. We now want to model  $p_j = \Pr(Y_j = 1)$  as a function of the linear expression  $X_{0j}\beta_0 + X_{1j}\beta_1 + \dots + X_{rj}\beta_r$ . There are three popular approaches for this problem. One is to let  $\hat{\beta}_0, \dots, \hat{\beta}_r$  be the usual estimate from a standard regression model. This is analogous to discriminant analysis where the variables  $X_{0j}, \dots, X_{rj}$  are not considered fixed known constants, but are themselves random variables (multivariate normal with constant covariance matrix) whose mean depends on the value of  $Y_j$ . The problem with this approach is that  $\hat{Y}_j = X_{0j}\hat{\beta}_0 + \dots + X_{rj}\hat{\beta}_r$  cannot be used directly to predict the value of  $p_j$ . Also, in many applications the  $X_{ij}$ 's are categorical, (e.g. province, occupation, etc.), thus violating the assumption of multivariate normality.

Two other popular approaches are known as probit analysis and logistic

regression. In probit analysis it is assumed that  $p_j = \Phi(\sum_i x_{ij}\beta_i)$ , where  $\Phi$  is the cumulative distribution function of a standard normal random variable. In logistic regression, it is assumed that

$$\theta_j = \log[p_j/(1 - p_j)] = \sum_i x_{ij}\beta_i.$$

Both these approaches are valuable analytic tools, and are available in many statistical packages (e.g. SAS, BMDP). The two approaches may be viewed together by letting

$$\eta_j = a(p_j) = \sum_i x_{ij}\beta_i.$$

For probit analysis we have  $\eta_j = \Phi^{-1}(p_j)$ , whereas for logistic regression we have  $\eta_j = \log [p_j/(1 - p_j)]$ . The maximum likelihood estimate for  $\beta_0, \dots, \beta_r$  is the solution to

$$\sum_j \frac{(y_j - \hat{p}_j)x_{ij}}{\hat{p}_j(1 - \hat{p}_j)q'(\hat{p}_j)} = 0, \quad \text{for } i = 0, \dots, r,$$

where  $a(\hat{p}_j) = \sum_i x_{ij}\hat{\beta}_i$ . These equations often must be solved iteratively. For the probit analysis we have

$$q'(p_j) = \frac{1}{\phi[\Phi^{-1}(p_j)]}$$

where  $\phi(\cdot)$  is the standard normal density function. For the logistic regression,

$$q'(p_j) = [p_j(1 - p_j)]^{-1}$$

so that the parameter estimate is given by the solution to

$$\sum_j (y_i - \hat{p}_j) X_{ij} = 0, \quad \text{for } i = 0, \dots, r.$$

The covariance matrix of  $\hat{\beta}_0, \dots, \hat{\beta}_r$  is  $A^{-1}$  where  $A$  is a matrix with  $(k, \ell)$ -th entry given by

$$A_{k\ell} = \sum_j \frac{X_{kj} X_{\ell j}}{p_j (1 - p_j) \{q'(p_j)\}^2}$$

This can be used to construct confidence intervals and perform hypothesis tests and model-building.

For logistic regression, the covariance simplifies to

$$A_{k\ell} = \sum_j p_j (1 - p_j) X_{kj} X_{\ell j}.$$

As an example of the utility of these models, we consider an unpublished analysis performed by Dolson and Morin on the Canadian Health and Disability Survey. The dependent variable was whether or not a person would be screened in as potentially disabled using the Screening Test 2 of the January 1983 Labour Force Supplement on Disability. For details, see Dolson and Morin (1983). Analysis was restricted to males aged 15-64. Of the 13,897 respondents, 14.4% (unweighted) were screened in. The screened-in rates are cross-classified by age-groupings, labour force participation and a proxy/non-proxy variable (with 3 levels: non-proxy, proxy by male or proxy by female) in Table 6. (The fitted values from the model to be discussed below are also shown.) The data are illustrated in Figure 5.

The fitted model reduced the number of parameters from 30 to 11. The final model was given by

$$\log[p_{ijk}/(1 - p_{ijk})] = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij},$$

where  $\sum \alpha_i = \sum \beta_j = \sum \gamma_k = 0$ ,  $\sum_j \delta_{ij} = 0$ ,  $\sum_i \delta_{ij} = 0$ , for the  $i$ -th age group,  $j$ -th labour force status and  $k$ -th proxy status (2 levels: non-proxy vs. proxy). The following were the estimated parameters.



<u>Parameter</u>	<u>Subscript</u>	<u>Estimate</u>
$\mu$		-1.43
$\alpha$	Age 15-24	-1.12
	Age 25-34	-0.571
	Age 35-44	0.0143
	Age 45-54	0.629
	Age 55-64	1.05
$\beta$	In Labour Force	-0.576
	Not in Labour Force	0.576
$\gamma$	Non-proxy	0.0859
	Proxy	-0.0859
$\delta$	Age 15-24, in L.F.	0.385
	Age 25-34, in L.F.	0.0938
	Age 35-44, in L.F.	-0.175
	Age 45-54, in L.F.	-0.243
	Age 55-64, in L.F.	-0.0612
	Age 15-24, not in L.F.	-0.385
	Age 25-34, not in L.F.	-0.0938
	Age 35-44, not in L.F.	0.175
	Age 45-54, not in L.F.	0.243
	Age 55-64, not in L.F.	0.0612

The fitted values are illustrated in Figure 6.

We see that even after adjusting for age and labour force status, there is a proxy effect on the screening rates. This proxy effect does not seem to depend on the sex of the proxy respondent. Also, there is no interaction between the proxy and the age/labour force status variables. This model does not necessarily imply a proxy bias, but it indicates that a proxy bias may potentially be present. Without a special study such as a re-interview program for the proxy respondent, it is impossible to definitively conclude the existence of a proxy bias.

## 4.2 Generalized Linear Models

In the previous section we discussed a large class of linear models related to the binomial model, of which probit analysis and logistic regression were special cases. We now extend these to the exponential family as proposed by Nelder and Wedderburn (1972).

As in Section 2.3, we assume  $y_j$  has probability function given by

$$f(y_j) = \exp[\kappa_j \{y_j \theta_j - b(\theta_j)\} + c(y_j, \kappa_j)],$$

where  $\mu_j = E[Y_j] = b'(\theta_j)$  and  $V_j = \text{Var}[Y_j] = b''(\theta_j)/\kappa_j$ .

We let  $\eta_j = q(\mu_j) = \sum_i X_{ij} \beta_i$  be the linear component of the model, where  $q(\cdot)$  is a known function.

Now the maximum likelihood estimates of  $\underline{\beta}$  are given by the solution to

$$\sum_j \frac{(y_j - \hat{\mu}_j) X_{ij}}{\hat{V}_j [q'(\hat{\mu}_j)]} = 0.$$

Nelder and Wedderburn (1972) have shown that a reasonable method for estimating  $\underline{\beta}$  is given by performing a number of weighted least-squares regressions, updating the weights and the dependent variables on successive iterations. This is called iteratively re-weighted least squares. In particular, the weights for the  $t$ -th iteration are given by

$$\hat{w}_j^{(t)} = \frac{1}{\hat{V}_j^{(t)} [q'(\hat{\mu}_j^{(t)})]^2}$$

and the dependent variables on the  $t$ -th iteration are given by

$$\hat{Z}_j^{(t)} = q(\hat{\mu}_j^{(t)}) + q'(\hat{\mu}_j^{(t)})(y_j - \hat{\mu}_j^{(t)}).$$

The  $(t + 1)$ -th iteration of  $\hat{\underline{\beta}}$  is then the solution to

$$\sum_j \hat{w}_j^{(t)} [\hat{Z}_j^{(t)} - \sum_{\ell} X_{\ell j} \hat{\beta}_{\ell}^{(t+1)}] X_{kj} = 0.$$

The estimated covariance matrix of  $\hat{\underline{\beta}}$  is given by  $A^{-1}$  where the  $(k, \ell)$ -th entry for  $A$  is

$$A_{k\ell} = \sum_j \hat{w}_j X_{kj} X_{\ell j}.$$

This implies that many standard weighted least-squares packages could be invoked to perform analysis of these generalized linear models.

For example, a common analysis of contingency tables, called log-linear models assumes a basic Poisson model with  $\log \mu_j = \sum_i X_{ij} \beta_i$ . Here we have

$$v_j = \mu_j,$$

$$q(\mu_j) = \log \mu_j,$$

so that the iteratively reweighted solution is given by assigning

$$\hat{w}_j^{(t)} = \hat{\mu}_j^{(t)},$$

$$\hat{z}_j^{(t)} = \log \hat{\mu}_j^{(t)} + \frac{y_j - \hat{\mu}_j^{(t)}}{\hat{\mu}_j^{(t)}}.$$

Hence, models similar to those described in Section 3 can be analyzed analogously using the generalized linear model formulation.

## 5. DIAGNOSTICS

Linear regression methods have been known now for over a century; see Hocking (1983) for a review of developments over the last 25 years. In more recent years attention has been focused on difficulties encountered when there is multicollinearity in the variables (leading to large variances of the parameter estimates) and when the models may fail. Some of these diagnostics are now available in SAS and SPSS-X.

The methods discussed in this paper extend linear regression to a much wider class of problems. Newer diagnostic techniques for models of this sort

are discussed in Landwehr, Pregibon and Shoemaker (1984).

In many statistical applications, the proposed model is only used as an approximation to reality. Therefore, the user of these models should employ these diagnostic tools in the course of the analysis.

## REFERENCES

- [1] Dolson, D. and Morin, J.-P. (1983). Disability data development project: Analysis of screening questionnaires. Technical Report, Health Division, Statistics Canada.
- [2] Hocking, R.R. (1983). Developments in linear regression methodology: 1959-1982 (with discussion). Technometrics, 25, pp. 219-249.
- [3] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Statist., 1, pp. 799-821.
- [4] Koch, G.G., Gilling, D.R., and Stokes, M.E. (1980). Biostatistical implications of design, sampling and measurement to health science data analysis. Ann. Rev. Public Health, 1, pp. 163-225.
- [5] Landwehr, J.M., Pregibon, D., and Shoemaker, A. (1984). Graphical methods for assessing logistic regression models (with discussion). J. Amer. Statist. Assoc., 79, pp. 61-83.
- [6] Little, R.J.A. (1982). Direct standardization: A tool for teaching linear models for unbalanced data. Amer. Statist., 36, pp. 38-43.
- [7] Nelder, J.A., and Wedderburn, R.W.M., (1972). Generalized linear models. J. Roy. Statist. Soc., Ser. A, 135, pp. 370-384.

**Table 1: Mean Number of Children Ever Born, by Marital Duration and Education Level. Sri Lanka 1975 (from Little, 1982)**

Years since First Marriage		Level of Education				
		No School	1 - 5 Years	6 - 9 Years	10+ Years	Row
0 - 4	Mean Count	0.96 112	0.88 376	0.95 442	0.92 351	0.92 1281
5 - 9	Mean Count	2.54 172	2.46 442	2.39 362	2.39 255	2.44 1231
10 - 14	Mean Count	3.87 197	3.91 482	3.73 293	3.14 145	3.76 1117
15 - 19	Mean Count	5.13 239	4.97 461	4.61 262	4.13 95	4.84 1057
20 - 24	Mean Count	6.22 292	5.87 377	5.22 184	4.47 40	5.79 993
25+	Mean Count	6.92 501	6.55 548	6.23 161	5.97 22	6.65 1232
Column	Mean Count	5.17 1513	4.24 2686	3.26 1704	2.30 908	3.94 6911



Table 2: Interactions for Mean Number of Children from Table 1

Years Since First Marriage		Level of Education				
		No School	1 - 5 Years	6 - 9 Years	10+ Years	Row
0 - 4	Raw Mean	0.96	0.88	0.95	0.92	0.92
	Adjusted Mean	1.31	1.07	0.86	0.71	1.02
	Interaction	-0.35	-0.19	0.09	0.21	
5 - 9	Raw Mean	2.54	2.46	2.39	2.39	2.44
	Adjusted Mean	2.78	2.54	2.33	2.18	2.49
	Interaction	-0.24	-0.08	0.06	0.21	
10 - 14	Raw Mean	3.87	3.91	3.73	3.14	3.76
	Adjusted Mean	4.06	3.82	3.61	3.46	3.77
	Interaction	-0.19	0.09	0.12	-0.32	
15 - 19	Raw Mean	5.13	4.97	4.61	4.13	4.84
	Adjusted Mean	5.11	4.87	4.66	4.51	4.82
	Interaction	0.02	0.10	-0.05	-0.38	
20 - 24	Raw Mean	6.22	5.87	5.22	4.47	5.79
	Adjusted Mean	6.01	5.77	5.56	5.41	5.72
	Interaction	0.21	0.10	-0.34	-0.94	
25+	Raw Mean	6.92	6.55	6.23	5.97	6.65
	Adjusted Mean	6.82	6.58	6.37	6.22	6.53
	Interaction	0.10	-0.03	-0.14	-0.25	
Column		5.17	4.24	3.26	2.30	3.94
		4.23	3.99	3.78	3.63	3.94

**Table 3: Analysis of Variance of Data from Table 1**

Source	Sum of Squares	Proportion of Total SS	DF	Mean Square	F	Signif. of F
Main Effects						
Marital Duration	27402.684	0.493	5	5480.537	1340.990	.000
Education/Duration	225.535	0.004	3	75.178	18.395	.000
Interactions						
Duration×Education	206.965	0.004	15	13.798	3.376	.000
Residual	27729.848	0.499	6787	4.986		
Total	55565.031		6810			

**Table 4: Physician Visits per Person per Year by Residence Size, Family Income and Education of Family Head, U.S. 1973**

Education in Years	Family Income		
	0 - 4999	5000 - 14999	15000 or more
SMSA			
Less than 12	6.15 (0.18)	4.73 (0.13)	4.82 (0.25)
12	6.17 (0.41)	4.98 (0.17)	4.70 (0.18)
More than 12	6.31 (0.49)	6.08 (0.19)	5.66 (0.16)
Non-SMSA			
Less than 12	5.08 (0.26)	4.14 (0.15)	4.42 (0.37)
12	5.36 (0.44)	4.32 (0.19)	4.49 (0.33)
More than 12	4.58 (0.58)	5.06 (0.29)	4.48 (0.31)

Note: Bracketed figures indicate standard errors of estimate.

Table 5. Estimated Physician Visits from Table 4.  
Original and Fitted Values

Education (in Years)		Family Income		
		0 - 4999	5000 - 14999	15000 or more
SMA				
Less than 12	Original	6.15 (0.18)	4.73 (0.13)	4.82 (0.25)
	Fitted	5.95 (0.07)	4.83 (0.07)	4.83 (0.07)
	Difference	0.20	-0.10	-0.01
12	Original	6.17 (0.41)	4.98 (0.17)	4.70 (0.18)
	Fitted	5.95 (0.07)	4.83 (0.07)	4.83 (0.07)
	Difference	0.22	0.15	-0.13
More than 12	Original	6.31 (0.49)	6.08 (0.19)	5.66 (0.16)
	Fitted	5.95 (0.07)	5.95 (0.07)	5.95 (0.07)
	Difference	0.36	0.13	-0.29
Non-SMSA				
Less than 12	Original	5.08 (0.26)	4.14 (0.15)	4.42 (0.37)
	Fitted	5.30 (0.11)	4.18 (0.11)	4.18 (0.11)
	Difference	-0.22	-0.04	0.24
12	Original	5.36 (0.44)	4.32 (0.19)	4.49 (0.33)
	Fitted	5.30 (0.11)	4.18 (0.11)	4.18 (0.11)
	Difference	0.06	0.14	0.31
More than 12	Original	4.58 (0.58)	5.06 (0.29)	4.48 (0.31)
	Fitted	5.30 (0.11)	5.30 (0.11)	5.30 (0.11)
	Difference	-0.72	-0.24	-0.82

**Table 6. Unadjusted and Fitted Screened-in Rates from Test 2.  
Canadian Health and Disability Survey, Males Aged 15-64,  
by Labour Force Participation and Proxy Status, Canada  
January 1983 (Unweighted)**

Age		Non-Proxy	Male Proxy	Female Proxy
In Labour Force				
15 - 24	Unadjusted	.065(.0067)	.055(.0143)	.056(.0069)
	Fitted	.065(.0051)	.056(.0044)	.056(.0044)
	Difference	.000	-.001	.000
25 - 34	Unadjusted	.085(.0058)	.058(.0252)	.069(.0069)
	Fitted	.085(.0048)	.071(.0046)	.071(.0046)
	Difference	.000	-.013	-.002
35 - 44	Unadjusted	.113(.0079)	.029(.0290)	.094(.0086)
	Fitted	.111(.0064)	.093(.0059)	.093(.0059)
	Difference	.002	-.064	.001
45 - 54	Unadjusted	.180(.0109)	.082(.0351)	.154(.0120)
	Fitted	.177(.0088)	.153(.0083)	.153(.0083)
	Difference	.003	-.071	.001
55 - 64	Unadjusted	.284(.0150)	.207(.0752)	.250(.0183)
	Fitted	.283(.0124)	.249(.0124)	.249(.0124)
	Difference	.001	-.042	.001
Not in Labour Force				
15 - 24	Unadjusted	.104(.0127)	.071(.0190)	.074(.0084)
	Fitted	.104(.0078)	.079(.0065)	.079(.0065)
	Difference	.000	-.008	-.005
25 - 34	Unadjusted	.146(.0239)	.367(.1450)	.227(.0365)
	Fitted	.192(.0213)	.167(.0194)	.167(.0194)
	Difference	-.046	.200	.060
35 - 44	Unadjusted	.348(.0372)	.455(.1501)	.324(.0544)
	Fitted	.359(.0309)	.320(.0299)	.320(.0299)
	Difference	-.011	.135	.004
45 - 54	Unadjusted	.534(.0361)	.625(.1712)	.454(.0505)
	Fitted	.525(.0293)	.483(.0301)	.483(.0301)
	Difference	.009	.142	-.029
55 - 64	Unadjusted	.571(.0220)	.563(.1240)	.591(.0420)
	Fitted	.585(.0194)	.543(.0217)	.543(.0217)
	Difference	-.014	.020	.048

NOTE: Bracketed figures are Standard Errors

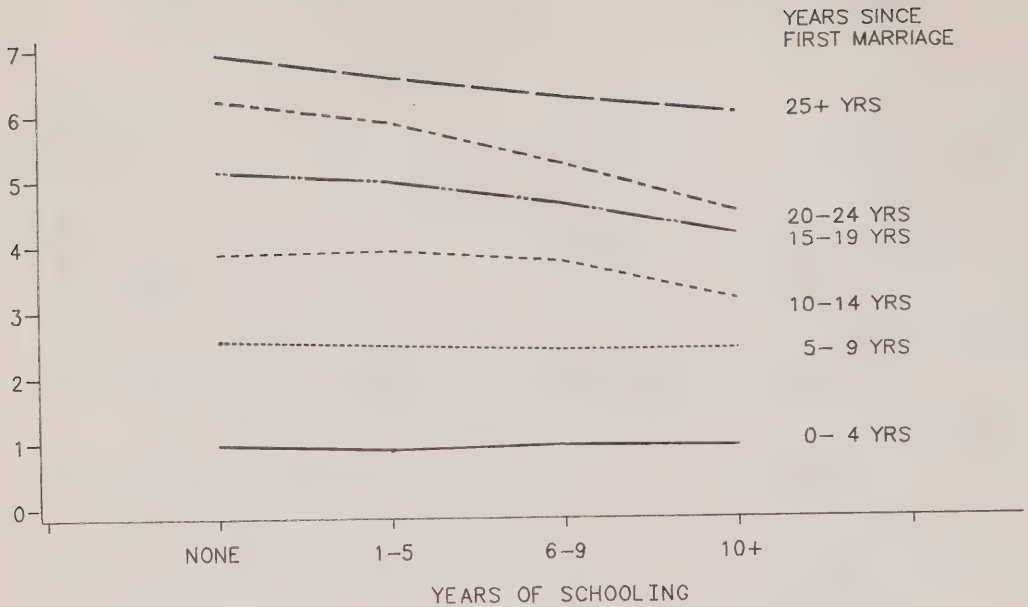


Figure 1: Observed Means from Sri Lanka Fertility Survey, 1975.  
Data source: Little (1982).

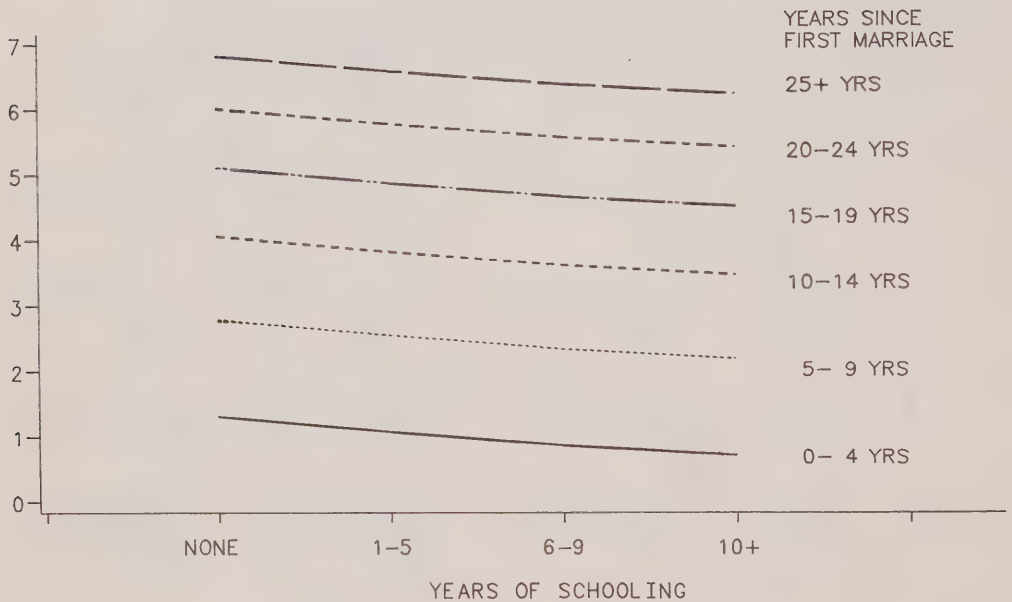


Figure 2: Adjusted Means from Sri Lanka Fertility Survey, 1975.  
Data source: Little (1982)



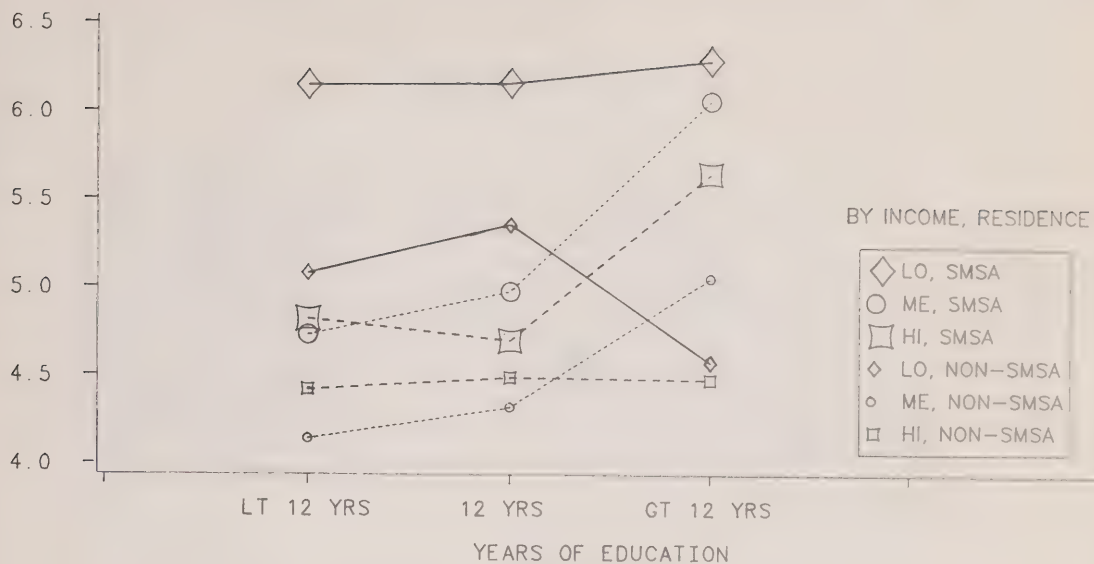


Figure 3: Observed Mean Number of Physician Visits per Person per Year. U.S.A., 1973.

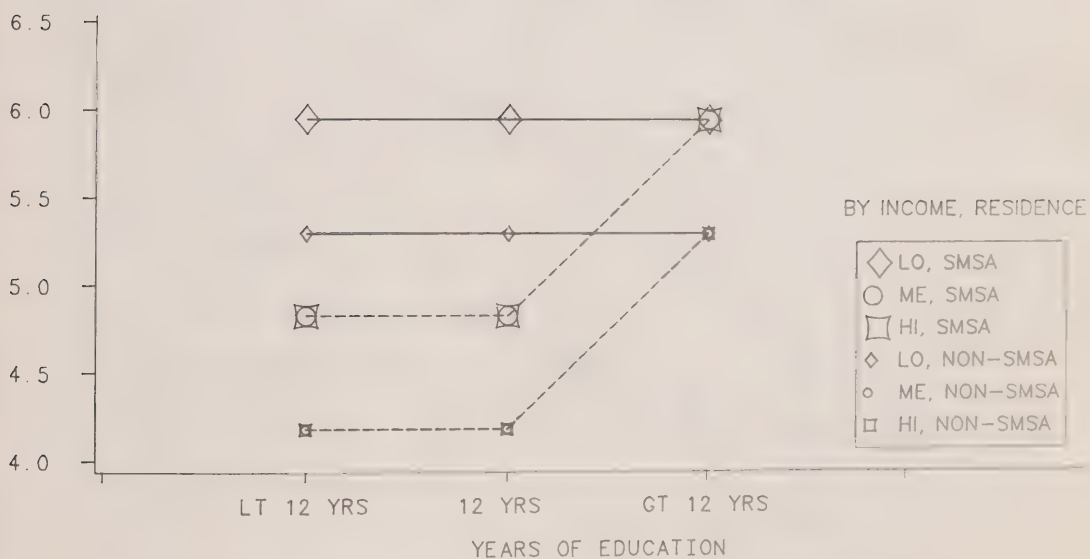


Figure 4: Model Predicted Mean Number of Physician Visits per Person per Year. U.S.A., 1973.

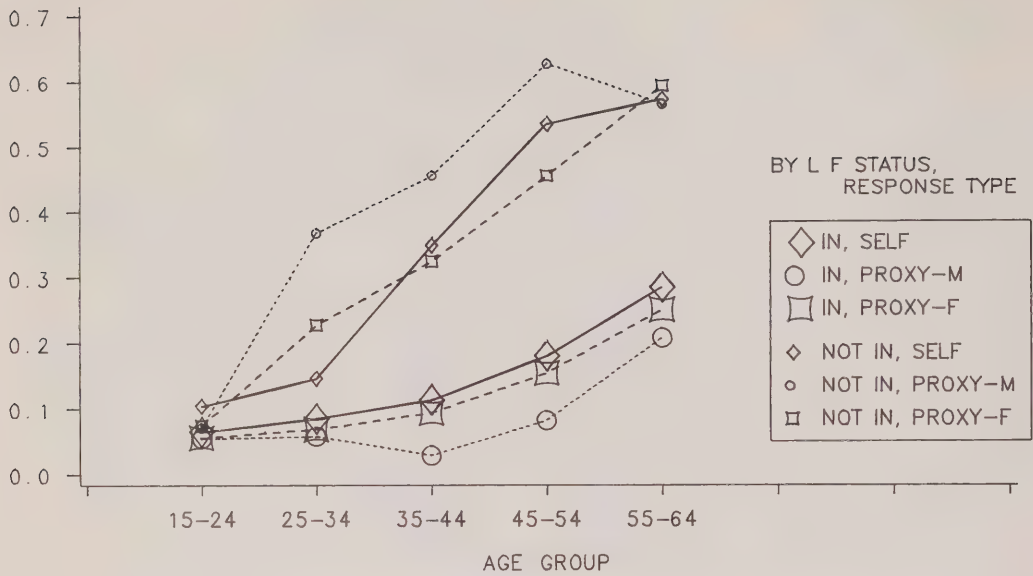


Figure 5: Observed Screening Rates, Disability Survey, January 1983, Males 15-64.

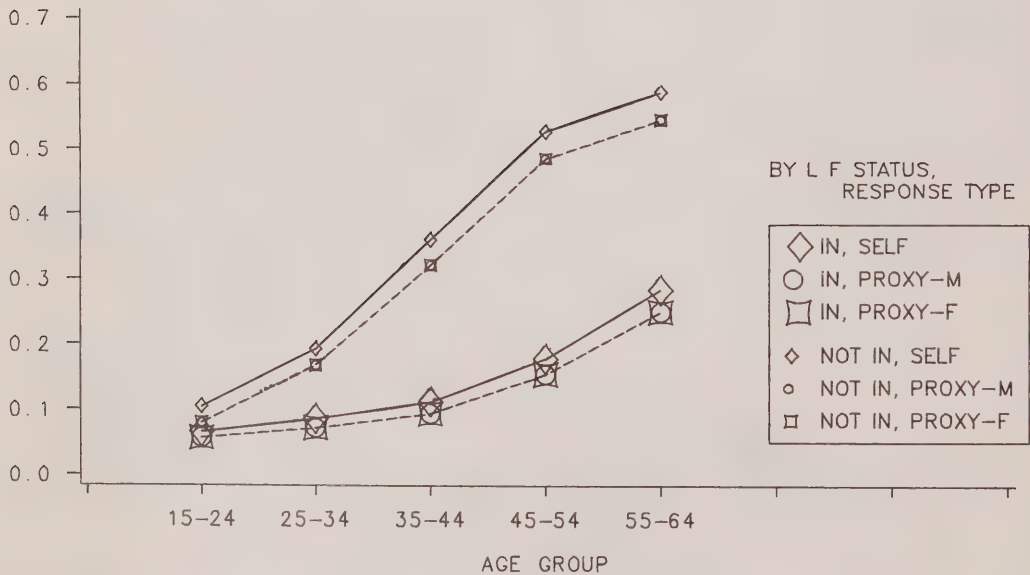


Figure 6: Predicted Screening Rates, Disability Survey, January 1983, Males 15-64.

## ADJUSTING SUB-ANNUAL SERIES TO YEARLY BENCHMARKS

Pierre A. Cholette<sup>1</sup>

This paper proposes a modification to the method of Denton (1971) for adjusting sub-annual series to yearly totals. These totals originate from more reliable sources and constitute annual benchmarks. The benchmarked series derived according to the modified method is more parallel to the unbenchmarked series than this is the case with the original method. An additive and a proportional variant of the method are presented. These can easily be adapted for flow, stock and index series. Also presented are a few recommendations about the preliminary benchmarking of current data and the management of "historical" estimates of the series.

## 1. INTRODUCTION

In many cases, the statistician obtains sub-annual data of a series from one source of data (such as a sample survey): and, the corresponding annual benchmark values from another more reliable source of data (such as a census). The annual sums of the observed sub-annual values are generally not equal to the annual benchmark values. Such sub-annual series require adjustment to annual benchmarks, that is benchmarking.

The solution proposed by Denton (1971) (and generalized by Fernandez in 1981) consists of finding a sub-annual series which would display the movement of the available sub-annual series as much as possible and whose annual sums (or averages) would match the more reliable annual benchmarks. The level of the resulting series would then be given by the annual benchmarks, whereas its movement would be dictated by the original sub-annual series. In other words, the adjusted or benchmarked series should run as parallel as possible to the original, while still satisfying the annual benchmarks. This paper suggests a modification to Denton's specification which makes the original and the adjusted series even more parallel.

We follow the model of Ehrenberg (1982) for the presentation of scientific

---

<sup>1</sup> Pierre A. Cholette, Time Series Research and Analysis, Statistics Canada.

papers. The reader will be exposed to the illustrations and results first; and the methodological details, afterwards.

## 2. ILLUSTRATION OF THE RESULTS

Figure 1 shows the corrections  $(x_t - z_t)$  made to the original series  $z_t$  according to the additive solution (with first differences) of Denton and according to the corresponding solution proposed in this paper. Since the corrections are to be added to the original sub-annual series  $z_t$ , the adjusted series  $x_t$  will be completely parallel to the original series, if and only if the corrections are constant. In the figure, this happens only for the corrections derived under the method proposed in this paper.

Figure 1 presented a trivial and ideal case which allowed the solution of constant corrections: All the average annual discrepancies, the differences between the annual benchmarks and the annual totals of the original series (divided by the number of months per year), were constant. Figure 2 displays a more realistic case, where the five average annual discrepancies vary about 200. As in the first example, the corrections derived by the herein proposed method are much more constant, especially in the first year.

As explained below, Denton's method does not only minimizes the change in the corrections (to make them as constant as possible) but also the size of the first correction. This can be seen both in Figures 1 and 2, where the first corrections are close to zero. The alternative solution, on the other hand, only minimizes the change in the corrections. Graphically this consists of fitting a curve through the average annual discrepancies, which is as flat as possible and which spans the same annual surfaces as the average annual discrepancies.

## 3. KEEPING THE ORIGINAL AND THE BENCHMARKED SERIES PARALLEL

Resuming the additive first difference formulation of Denton as well as his notation, the desired series  $x_t$  minimizes the following objective function

$$p(x) = \sum_{t=1}^n (\Delta x_t - \Delta z_t)^2 = \sum_{t=1}^n (\Delta(x_t - z_t))^2, \quad x_0 = z_0, \quad (1)$$

where  $z_t$  stands for the original sub-annual series at time  $t$ . This function is minimized subject to the equality constraints between the annual sums of the values obtained and the available benchmarks  $y_i$ :

$$\sum_{t=(i-1)k+1}^{ik} x_t = y_i, \quad i = 1, 2, \dots, m. \quad (2)$$

where  $k$  is the number of "months" per year.

Denton justifies hypothesis  $x_0 = z_0$  claiming that it is legitimate to assume the equality of the last fitted and observed values prior to the estimation interval. Objective function (1) would then mean that the adjusted series  $x_t$  should have the same slope as the original series  $z_t$ ; and therefore, that the slope of the differences between the two series should be minimized (subject to the constraints). However, after substituting  $x_0 = z_0$ , objective function (1) can be rewritten as:

$$p(x) = (x_1 - z_1)^2 + \sum_{t=2}^n (\Delta(x_t - z_t))^2. \quad (3)$$

This transformation emphasizes that the assumption  $x_0 = z_0$  implies minimizing the size of the first correction. As illustrated in Figures 1 and 2, minimizing the first correction pulls the correction curve towards zero at the start of the series. This produces a wave in the first year which is transmitted to the other years. This wave in the corrections prevents, by definition, the maximum parallelism between the observed and adjusted series.

The specification proposed here simply refrains from postulating  $x_0 = z_0$  and yields the following objective function

$$p(x) = \sum_{t=2}^n (\Delta(x_t - z_t))^2, \quad (4)$$

subject to the same constraints of equation (2).

In linear algebra, the constrained objective function is written

$$u(\underline{x}, \underline{q}) = (\underline{x} - \underline{z})' \underline{A} (\underline{x} - \underline{z}) - 2 \underline{q}' (\underline{y} - \underline{B}' \underline{x}), \quad (5)$$



where the vectors and matrices involved are:

$$\underline{x}_{n \times 1} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \underline{z}_{n \times 1} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}, \quad \underline{y}_{m \times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad \underline{q}_{m \times 1} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{bmatrix}, \quad (6)$$

$$\underline{A}_{n \times n} = \underline{D}' \underline{D}, \quad (\underline{n}-1) \times n = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}, \quad (7)$$

$$\underline{B}_{n \times m} = \begin{bmatrix} \frac{j}{n} & 0 & \dots \\ \vdots & \frac{j}{n} & \dots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{bmatrix}, \quad \underline{j}_{k \times 1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad (n = km). \quad (8)$$

Vector  $\underline{q}$  contains the Lagrangian multipliers. Variables  $n (= mk)$ ,  $m$  and  $k$  respectively stand for the number of observations and of years in the series and the number of months per year.

The normal equations associated with objective function (5) are

$$\begin{aligned} \underline{du}/\underline{dx} &= (\underline{A} + \underline{A}')(\underline{x} - \underline{z}) + 2 \underline{B} \underline{q} = \underline{n} \\ \underline{du}/\underline{dq} &= 2(\underline{B}' \underline{x} - \underline{y}) = \underline{n} \end{aligned} \quad (9)$$

and yield solution

$$\begin{bmatrix} \underline{x} \\ \underline{q} \end{bmatrix} = \begin{bmatrix} \underline{A} & \underline{B} \\ \underline{B}' & \underline{0} \end{bmatrix}^{-1} \begin{bmatrix} \underline{A} & \underline{0} \\ \underline{0} & \underline{I} \end{bmatrix} \begin{bmatrix} \underline{z} \\ \underline{y} \end{bmatrix} = \frac{W}{(n+m) \times (n+m)} \begin{bmatrix} \underline{z} \\ \underline{y} \end{bmatrix}. \quad (10)$$

Substituting identity  $y = B'z + r$ , where  $r$  contains the  $m$  annual discrepancies, gives

$$\begin{bmatrix} \underline{x} \\ \underline{a} \end{bmatrix} = \begin{bmatrix} \underline{A} & \underline{B} \\ \underline{B}' & \underline{0} \end{bmatrix}^{-1} \begin{bmatrix} \underline{A} & \underline{0} \\ \underline{B}' & \underline{I} \end{bmatrix} \begin{bmatrix} \underline{z} \\ \underline{r} \end{bmatrix} = \begin{bmatrix} \underline{I} & \underline{W}_x \\ \underline{0} & \underline{W}_1 \end{bmatrix} \begin{bmatrix} \underline{z} \\ \underline{r} \end{bmatrix} \Rightarrow \underline{x} = \underline{z} + \underline{W}_x \underline{r}. \quad (11)$$

This reformulation of the solution reduces computing time in the application of the calculated weights compared to formulation (10). Also note that once the weights  $\underline{W}_x$  are obtained, they can be used for any number of series having the same number of observations. Furthermore, we recommend (Choi, 1978, section 6; 1979, 4.3) to compute  $\underline{W}_x$  for a 5-year interval and to use it in a moving average manner (moving one year at the time) for series of 5 years and more. Apart from saving on calculations, this procedure generates only two revisions in the estimates (*ceteris paribus*) when new years of observations are added to the series.

Denton solves the inversion in equation (10) by parts. This is impossible here since matrix  $\underline{A}$  is singular. The overall matrix however is not singular and can be inverted.

In fact, the method developed herein uses the solution proposed by Root, Feibes and Lisman (1967) to interpolate between annual data in the absence of sub-annual information. Solution (11) exactly consists in interpolating between the annual discrepancies with the method of these authors and in adding the resulting estimates (the corrections) to the original sub-annual series.

#### 4. PROPORTIONAL VARIANT

The proportional method now presented in this section is also a variant of Denton's proportional method, from which  $x_0 = z_0$  was removed. As in Section 2, the objective function still minimizes the sum of the squared differences between the slopes of the original and desired sub-annual series  $\{z_t\}$  and  $\{x_t\}$ . Each term in the sum is weighted however by the value of the corresponding sub-annual observation:

$$p(x) = \sum_{t=2}^n (\Delta(x_t - z_t)/z_t)^2 = \sum_{t=2}^n (\Delta(x_t/z_t))^2. \quad (12)$$

This variant is suitable for series with strong seasonality, when it is thought that seasonal trough months cannot account for the annual discrepancy as much as seasonal peak months: The size of the corrections are proportional to the level of each observation, as illustrated in Figure 3. The low observations get smaller corrections than the seasonally higher observations, although the minimized proportional corrections  $x_t/z_t$  are as flat as permitted by the annual discrepancies. Note that with the proportional variant all observations must be positive and that all the adjusted values will also be positive.

It can also be shown (Cholette, 1978, Section 3; 1979, 3) that the proportional variant is a linear approximation of the strongly non-linear growth rate preservation method (Smith, 1977; Helfand et al., 1978), which would have the following objective function:

$$p(x) = \sum_{t=2}^n (x_t/x_{t-1} - z_t/z_{t-1})^2. \quad (13)$$

The approximation is exact in situations of constant annual proportional discrepancies on the estimation interval.

In linear algebra, the constrained objective function associated to the proportional method is

$$\underline{u}(\underline{x}, \underline{q}) = (\underline{x} - \underline{z})' \underline{Z}^{-1} \underline{A} \underline{Z}^{-1} (\underline{x} - \underline{z}) - 2 \underline{q}' (\underline{y} - \underline{B}' \underline{x}), \quad (14)$$

where  $\underline{Z}^{-1}$  is a diagonal matrix with elements  $1/z_1, 1/z_2, \dots$ . The solution has the same structure as the additive variant ( $\underline{Z}^{-1} \underline{A} \underline{Z}^{-1}$  replacing  $\underline{A}$  in (11)) and writes:

$$\begin{bmatrix} \underline{x} \\ \underline{q} \end{bmatrix} = \begin{bmatrix} \underline{Z}^{-1} \underline{A} \underline{Z}^{-1} & \underline{B}' \\ \underline{B}' & \underline{0} \end{bmatrix}^{-1} \begin{bmatrix} \underline{Z}^{-1} \underline{A} \underline{Z}^{-1} & \underline{0} \\ \underline{B}' & \underline{I} \end{bmatrix} \begin{bmatrix} \underline{z} \\ \underline{r} \end{bmatrix} = \begin{bmatrix} \underline{I} & \underline{W}_x \\ \underline{0} & \underline{W}_1 \end{bmatrix} \begin{bmatrix} \underline{z} \\ \underline{r} \end{bmatrix}. \quad (15)$$

Unlike the weights in the additive variant however, weights  $\underline{W}_x$  of the proportional solution must be computed for each series and even for each

application interval of a given series.

## 5. STOCK AND INDEX SERIES

The additive and proportional variants of the method presented above are designed for flow series, whose annual values correspond to the sum of the sub-annual values. The solutions can very easily be adapted for stock series, whose annual values are associated to only one sub-annual value (usually that of the last month); and for index series, whose annual values correspond to the average of the sub-annual values. For a quarterly stock series, for instance, one merely has to redefine the component vector  $\underline{j}$  of matrix  $\underline{B}$  as

$$\frac{\underline{j}'}{1 \times 4} = [0 \ 0 \ 0 \ 1];$$

and, for monthly index series as

$$\frac{\underline{j}'}{1 \times 12} = [1/12 \ 1/12 \ \dots \ 1/12].$$

## 6. DISCUSSION

### 6.1 Historical Data

There is a lot of confusion regarding the interpretation of assumption  $x_0 = z_0$  of Denton. In that respect, the author writes: "It is assumed that no adjustments are to be made to the original series for years outside the range from year 1 to  $m$ , inclusive." (p. 100, above equation (3.2)).

If these years are left untouched because they never had any benchmarks, the solution proposed by Denton is defensible: No corrections result for years -1 and 0; and small and gradually introduced corrections, at the start of year 1. (Remember that  $x_0 = z_0$  implies minimizing the first correction.) The resulting adjusted series is then continuous as illustrated in Figure 4, curve ADER.

However, if the first years are left untouched because they were already

benchmarked and are now considered "historical", we do not agree with assumption  $x_0 = z_0$ . Indeed, this assumption will generally produce a discontinuity between years 0 and 1, as shown in Figure 4 by curve A'CDEB. Years -1 and 0 have already received corrections of magnitude around CD, whereas the start of year 1 receives corrections which are as small as possible.

In order to "freeze" the historical data after a certain number of years, two solutions are possible. First, one can explicitly specify the freezing constraint in the objective function which becomes

$$p(x) = ((x_1 - z_1) - (x_0 - z_0))^2 + \sum_{t=2}^n (\Delta(x_t - z_t))^2, \quad (16)$$

where  $(x_0 - z_0)$  is known and equal to the last correction used for historical year 0. This correction is generally not equal to zero (Cholette, 1979b, 1983). This specification amounts to determining the starting point of the correction curve.

Second, a less specific but equally effective solution consists of applying the methodology already proposed in this paper (additive or proportional versions) as a moving average, which moves one year at the time. With a 5-year estimation interval, for instance, the estimates automatically become final after two years of revision; and, after one year, in the case of a 3-year interval (Cholette, 1978, section 6 a; 1979, 4.3). The resulting benchmarked series is continuous, as illustrated in Figure 4 by curve A'CB.

## 6.2 Implementation

The practitioners of benchmarking have a tendency to feed to the benchmarking programme the already benchmarked years of data followed by one year of unbenchmarked data (all accompanied by their benchmarks). For methodologists, it is obvious that one must always submit the unbenchmarked data (with the yearly benchmarks). Feeding benchmarked data will generally induce an artificial seasonal movement in the resulting benchmarked series (Cholette, 1978, Section 6b).



### 6.3 Preliminary Benchmarking of Current Data

A final comment is in order. During a current (uncompleted) year, one cannot calculate growth rates, for instance, between the benchmarked segment of the series (AB) and the unbenchmarked segment (CD). Doing so usually produces a discontinuity BC between the two segments AB and CD as illustrated in Figure 5 by curve ABCD.

Two solutions are then possible. One, the inter-temporal comparisons are based only on the unbenchmarked data. Two, the current data are preliminarily benchmarked by repeating the last available correction BC for the current year. (Note that including the incomplete current year in the objective function (4) (or 12) would yield identical preliminarily benchmarked values.) One can then compare the benchmarked segment AB with the preliminarily benchmarked segment BE as illustrated in Figure 5 by curve ABE. We favour this second alternative.

### 6.4 Relation with Other Methods

The Denton (1971) benchmarking method, the modified Denton method (presented in this paper), the methods of Glejser (1966), of Boot, Feibes and Lisman (1967), of Lisman and Sandee (1964), and of Bassie (1939) could be referred to as univariate methods. No series other than that considered and its annual benchmarks enter the benchmarking process. On the contrary, the methods by Friedman (1962), by Chow and Lin (1971), by Somermeyer, Jansen and Louter (1976) and by Wilcox (1983) are multivariate. Auxiliary series are used in the computation of the desired series.

For instance, Chow and Lin (1971) proposed a method to obtain the desired sub-annual series from yearly totals and from related series. The movement of the resulting series is as much as possible similar to the movements of the related series (and the series obtained satisfies the annual constraints). Fernandez (1981) observes that the Chow and Lin method can produce movement discontinuities between the years. He then proposes a synthesis of the Chow-Lin and of the Denton methods. The combined method eliminates the inter-annual discontinuities, but still relies on the hypothesis  $x_0 = z_0$ . As illustrated above, this hypothesis often introduces spurious fluctuations in the calculated series. We would think that it should be possible to refrain from the hypothesis in the case of Fernandez as in the case of Denton.

## 7. SUMMARY AND CONCLUSIONS

Denton (1971) intended to keep the original and benchmarked series as parallel as made possible by the annual discrepancies. This paper suggested a modification to the benchmarking method which makes the original and benchmarked series more parallel than is the case with the original method. This improvement holds both for the additive and the proportional variants of the method. We suspect that the generalized multivariate method by Fernandez could be improved in the same direction.

The method proposed can very easily be adapted for flow, stock as well as index series.

Before making intertemporal comparisons between the benchmarked and current data, it is essential to preliminarily benchmark the current data (in the manner proposed).

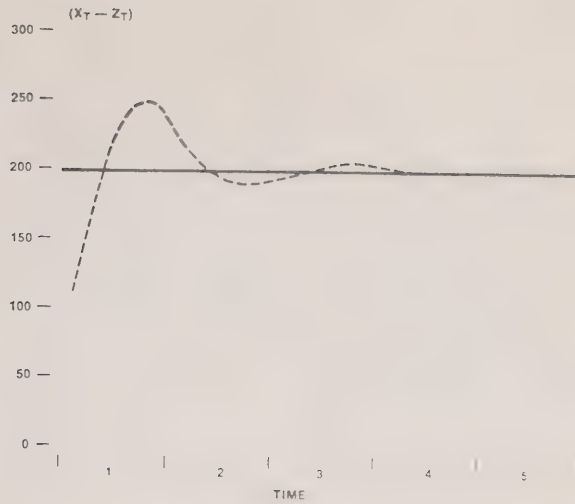
The suggested 5-year moving average implementation of the method will automatically "freeze" the past estimates after two years of revision.

## REFERENCES

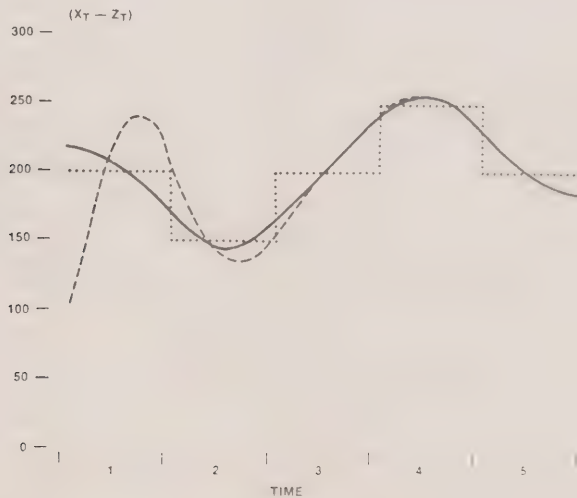
- [1] Baldwin, A. (1978), "New Benchmarking Algorithms using Quadratic Minimization." National Product Division, Statistics Canada, Research Paper.
- [2] Rassie, B.L. (1939), "Interpolation Formulae for the Adjustment of Index Numbers," Proceedings of the Annual Meetings of the American Statistical Association
- [3] Boot, J.C.G., Feibes, W., Lisman, J.H.C. (1967), "Further Methods of Derivation of Quarterly Figures from Annual Data," Applied Statistics, Vol. 16, No. 1, pp.65-75.
- [4] Cholette, P.A. (1978), "A Comparison and Assessment Various Adjustment Methods of Sub-Annual Series to Yearly Benchmarks," Time Series Research and Analysis, Statistics Canada, Research Paper 78-03-001R.

- [5] Cholette, P.A. (1979a), "Adjustment Methods of Sub-Annual Series to Yearly Benchmarks," Proceedings of the Computer Science and Statistics, 12th Annual Symposium on the Interface, J.F. Gentleman Ed., University of Waterloo, pp. 358-36.
- [6] Cholette, P.A. (1979b), "A Note on 'Freezing' Past Estimates when Benchmarking," Time Series Research and Analysis, Statistics Canada, Research Paper 79-06-002E.
- [7] Cholette, P.A. (1982), "Minimum Quadratic Adjustment Program (MQAP-I) of Series to Annual Totals - Users Manual," Time Series Research and Analysis, Statistics Canada, 82-11-003R.
- [8] Cholette, P.A. (1983), "Benchmarking Series with Bi-Annual Benchmarks when Knowing the Ending Point," Time Series Research and Analysis, Statistics Canada, Research Paper 83-05-002B.
- [9] Chow, G.C., Lin, An-loh (1971), "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series," Review of Economics and Statistics, Vol. 53, No. 4, pp. 372-375.
- [10] Dagum, E.B. (1977), "Comparison of Various Interpolation Procedures for Benchmarking Economic Time Series," Time Series Research and Analysis, Statistics Canada, Research Paper 77-05-006E.
- [11] Denton, F.T. (1971), "Adjustment on Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization," J.A.S.A., Vol. 66, No. 333, pp. 99-102.
- [12] Fernandez, R.B. (1981), "A Methodological Note on the Estimation of Time Series," Review of Economic and Statistics, Vol. 63, pp. 471-476.
- [13] Friedman, M. (1962), "The Interpolation of Time Series by Related Series," J.A.S.A., Vol. 57, No. 300, pp. 729-757.

- [14] Glejser, H. (1966), "Une méthode d'évaluation de données mensuelles à partir d'indices trimestriels ou annuels," Cahiers Economiques de Bruxelles, No. 19, 1er trimestre, pp. 45-64.
- [15] Helfand, S.D. Monsour, N.J, Trager, M.L. (1978), "Historical Revision of Current Business Survey Estimates," U.S. Bureau of the Census, (Research Paper).
- [16] Huot, G. (1975), "Quadratic Minimization of Monthly Estimates to Annual Totals," Time Series Research and Analysis, Statistics Canada, Research Paper 75-11 M10E.
- [17] Lisman, J.H.C., Sandee, J. (1964), "Derivation of Quarterly Figures from Annual Data," Applied Statistics, Vol. 13, No. 2, pp. 87-90.
- [18] Smith, P. (1977), "Alternative Method for Step Adjustment," Current Economic Analysis Division, Statistics Canada, Research Paper.
- [19] Somermeyer, W.H, Jansen, R., Lauter, A.S. (1976), "Estimating Quarterly Values from Annually Known Variables in Quarterly Relationships," J.A.S.A, Vol. 71, No 355, pp. 588-595.
- [20] Wilcox, J.A. (1983), "Disaggregating Data Using Related Series," Journal of Business and Economic Statistics, Vol. 1, No 3, pp. 187-191.

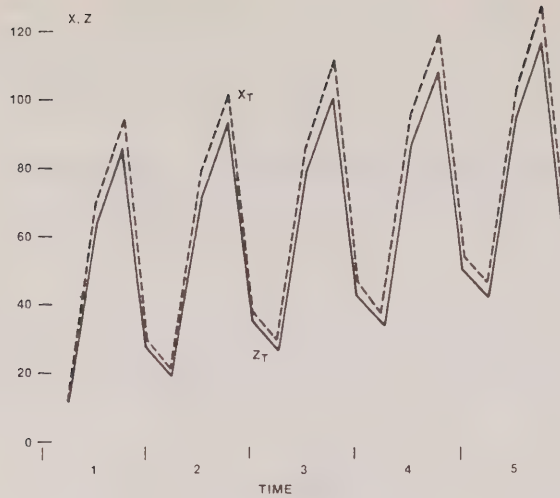


**Figure 1:** Corrections  $(x_t - z_t)$  made to the unbenchmarked series according to Denton's method (dashed line) and according to the method proposed in this paper (solid) in an ideal situation of constant annual discrepancies.



**Figure 2:** Corrections  $(x_t - z_t)$  made to the unbenchmarked series according to Denton's method (dashed line) and according to the benchmarking method proposed in this paper (solid) in a situation of variable average annual discrepancies (dotted).

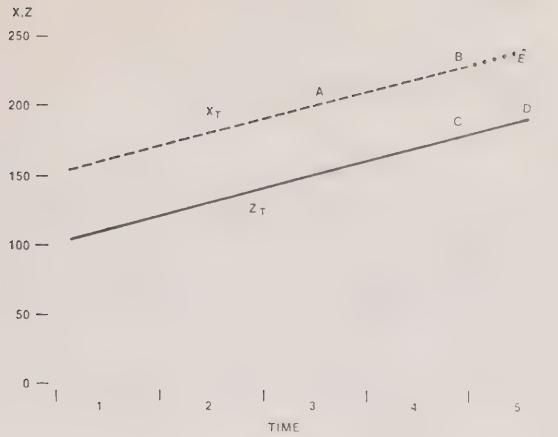




**Figure 3:** Original series (solid curve) and benchmarked series (dashed) according to the proportional variant of the benchmarking method proposed in this paper (in a situation of constant annual proportional discrepancies).



**Figure 4:** Benchmarked series according to Denton's method, when there are no benchmarks for year -1 and 0 (curve ADEB) and when there are benchmarks and year -1 and 0 were already benchmarked (A'CDEB): and according to the method proposed in this paper, applied in a moving average manner, when there are benchmarks for years -1 and 0 (A'B).



**Figure 5:** Continuity between the benchmarked series (dashed curve) and the preliminarily benchmarked series (dotted) and discontinuity BC between the benchmarked (dashed) and the unbenchmarkd (solid) series.

## EXAMINING EXPENDITURES ON ENERGY

Louise A. Heslop<sup>1</sup>

Using data from the Family Expenditures Surveys over time, consumer expenditures on in-home and transportation energy from 1969 to 1982 are being studied. This article briefly summarizes some of the procedures being used to explore the data, summarize it and develop insights into shifts in consumption for policy implications purposes. With such a complex data set and such a complex, multi-faceted subject for analysis some effort must be made to reduce information flows and at the same time increase the information content of each factor of both input and output in the analyses.

## 1. THE ENERGY ISSUE

To some, energy conservation may be a dead issue. There is no shortage of energy (maybe never was): prices for energy have stabilized.

Energy matters dominated the 1970's having major impacts on the world economic order and creating international strife. Domestically they impacted drastically on federal - provincial relations and business - government relations and on family budgets: caused the restructuring of the manufacturing base, the auto industry, etc. Despite its reported demise as an important issue, energy consumption and prices remain as high priority concerns of consumers, businesses and governments. Energy conservation has lost its sparkle but not its real value.

The research I will be reporting on briefly has been developed in consultation with policy makers in Consumer and Corporate Affairs Canada and Energy, Mines and Resources Canada which continue to run active research programmes on consumer energy use and conservation. The project structure has taken their interests, orientations and limitations into consideration.

Also, within the last five years an international group of social scientists has begun a series of research and information exchanges on consumer behaviour and energy use. As a member of that group I have been keenly aware

---

<sup>1</sup> Louise A. Heslop, Research and Analysis Division, Statistics Canada.

of the problems and prospects and the current state of knowledge and research techniques of that group.

## 2. PROBLEMS IN ENERGY RESEARCH

Perhaps the major problem in studies of consumer energy use has been to obtain reasonably reliable measures of use from sufficiently large and representative samples. Getting such data over a period of time, especially a time period spanning the infamous 1973 oil embargo period, would send a researcher into Nirvana. The Family Expenditure data collected by the Consumer Income and Expenditure Division of Statistics Canada come close enough to these requirements to at least set one's heart fluttering. It is a series of retrospective recall studies conducted for the years 1969, 1972, 1974, 1976, 1978 and 1982. So it covers the time period of interest for a large sample and the sampling technique used ensures that the design is representative of Canada for those areas studied, usually urban centres. Additionally it contains a great many other variables of interest in any study of energy use, e.g., home ownership, some house characteristics, vehicle and appliance ownership, family characteristics and expenditures on other categories of consumer goods and services, etc.

Most studies which attempt a measure of consumer expenditures rely on recall or file checking by respondents. There are obvious problems with the accuracy of such data on an individual basis. The problems are less restrictive with very large samples. For most independent studies, the costs of such large samples are prohibitive. However, FAMEX sample sizes are very large.

Only one major study in Canada has used independent record checking, obtaining records from suppliers by household with the permission of the household, but through this technique was able to obtain electricity use records on less than half of its sample. Natural gas and oil records were obtained on only about one-third of the sample. This procedure of record checking is highly accurate, removes the problems associated with recall, especially over long periods of time, and of reporting bias of respondents. However, practically it is impossible to use for large samples across the country.

Although the FAMEX Study uses recall procedures, the information on energy

expenditures are not likely to be as biased as in a study specifically designed to record energy behaviours since respondents are not sensitized to the subject of the study. Also the data from pre-energy crisis periods was collected in the same way as that since the crisis, again reducing the likelihood of response bias. So the FAMEX data set offers a unique opportunity to examine a very large set of samples during a very important period of time.

The data set is not without its problems, some because of the sampling procedure and some because of the inherent complexity of any study of energy use. Changes in expenditure categories and their contents, especially those other than energy, have required that we manipulate the data considerably to create consistency across years. It is not possible to track in-home energy expenditures for those families who do not pay for energy directly, i.e., apartment dwellers with central metering and roomers. Some researchers have imputed values to these households based on their rents but we chose not to, and instead have chosen to restrict our study to those households who have the ability to monitor and affect their own energy use. These households are the consumer groups who will be the focus of any government programmes to alter consumer consumption.

There are several factors which make the study and the altering of energy consumption of households difficult:

- Capital commitments restrict the ability of the household to respond in the short-term and increase the cost of response - e.g., house size, number and type of appliances, size and number of vehicles. Some studies have noted that home characteristics alone may account for 24% of in-home energy consumption. Family size may be considered as a capital commitment as well.
- Flow feasibilities - There are restrictions in the ability to change the amount and types of fuels used depending on the technology and fuels available under different circumstances and for varying amounts of money, e.g., natural gas heating is not available to rural residents: instantaneous changes can not be made in the type of home heating fuel used.
- Exogenous factors affect the amount of energy needed for similar



performance in different situations, e.g., weather, distances between points in cities, etc.

### 3. SUMMARIZING INFORMATION INPUTS AND MAXIMIZING INFORMATION OUTPUTS

With such a complex data set and such a complex, multi-faceted subject for analysis some effort must be taken to reduce information flows and at the same time increase the information content of each factor of both input and output. There are several ways of doing this, some of which we will be using, they include:

a) Constructing Complex Input Variables - to reduce the number of factors being studied to the most salient ones.

i) Discontinuous complex input variables were created by combining income and transportation energy consumption but not as continuous variables. Rather groupings were created to develop a set of typologies whose characteristics can then be examined for differences. In this case the groupings were developed by creating expenditure quartiles for each energy category, collapsing the two middle categories, and then combining the two resulting three cells into a nine cell matrix of interrelated categories (see Table 1, source: McDougall, Ritchie and Claxton). In particular, the corner cells are of interest in contrast to each other and to the middle cell. This typology was developed in an earlier study for Consumer and Corporate Affairs Canada. So comparing the output from the FAMEX data to the data set used in the CCA study will be of particular interest. Comparing the characteristics of these groupings over time will also be of interest. For example, do the Churchmice continue to be impoverished Canadians (involuntary simplicity) or is there any indication that there is some voluntary embracing of low energy, lifestyles? In Table 2 the characteristics of three cells of the typology from two different years are compared - the Churchmice, the Roadrunners and the Hippos. Looking first at the Churchmice, information on a selection of possible analysis variables is shown across two different years, 1974 and 1978. To simplify for this presentation only the rankings of the cell within the typology set of cells is given. Characteristically those consuming the least amount of energy

have had the least resources in general, i.e., the lowest incomes, the lowest levels of education, the oldest. These characteristics are evident for the Churchmice in 1974, they also have the lowest levels of consumption for all the expenditure categories shown. Although they are the oldest group they do not have the lowest number of very young children. Probably this group consists of a mix of senior citizens and single parent households (probably headed by women) with young children. Note that this group also has the lowest number of full-time earners (F-T earners). In 1978 the general picture is still the same except that this group is no longer the oldest. In fact the oldest group is in the adjacent cell to the right in the typology (not shown here). It would seem that in 1978 the very old are consuming a relatively larger amount of in-home energy. Perhaps this group is financially better off in 1978 than in 1974 or perhaps they have been unable to hold the line on energy expenditures as prices have risen.

In 1974 the Hippos also fit expectations. They seem to be middle-aged with large numbers of children 5-16 years of age. The "full nest" family, they spend the largest amount on most expenditure categories. They are also the most highly educated. In 1978 this is no longer true as the education ranking of this cell has dropped. Also this group no longer has the highest shelter expenditure. Some suggestions for these observations may be that those with the largest homes and the highest education have begun to modify their homes to reduce energy expenditures.

The Roadrunners have changed also. In 1974 they were the youngest group with very small families. In 1978 they appear to be characterized as young families with young children. One of the most dramatic changes for this group has been that their alcoholic beverages and tobacco expenditures have dropped dramatically.

The significance of these changes can be determined with appropriate statistical tests. The purpose of this discussion was to introduce the idea of searching for meaningful typologies within the data. Pictures of the lifestyles of the groups emerge which can be very useful in furthering conservation programmes directed at each group.

Further analysis may look not at level of expenditures but at percent of expenditures. Such an analysis will reveal the characteristics of those who are most heavily burdened with energy bills.

ii) Continuous complex input variables can be constructed to eliminate the effects of variables known to have very large effects, but ones which are difficult or impossible for consumers to manage.

In-home energy expenditures can be examined for factors related to them, but since one of the main determinants of in-home energy expenditures is house size, this size factor can be absorbed into the input variable to allow for examination of other more relevant (from a policy perspective) factors. So instead of in-home energy expenditures, in-home expenditures/room are examined. Taking this one step further, climate and weather variations from year to year may be controlled for by looking at expenditures/room/degree day. This last factor is added to the data set by city by year. Degree day data for each year for each city were obtained from Environment Canada. Table 3 indicates how the figures change as the factor studied becomes more complex again across two of the years of data. A comparison of the two years and differences in the measures of change between years suggests the importance of refining the measure to improve understanding of the process.

b) Constructing summary output variables to examine the structure of the data - Example of regression coefficients.

In Tables 4-6 some regression outputs are presented. Three models are examined. In each succeeding model the dependent variable becomes more complex. In so doing the factors known to impact significantly on energy consumption can be controlled for and the effects of the remaining variables examined more constructively for any significant explanatory power.

In these analyses no attempt has been made to deal with the problem of the complex sampling design. A future analysis will do so using the Taylor linearization procedure and results will be compared. However, the results from both a weighted and an unweighted sample are shown for 1974. As can be seen the values of the coefficients change very little and their significance or lack thereof does not change. Because of the restrictions indicated and also the fact that the very large sample sizes are used here produce significant results under conditions of very slight differences, it is advised that great care be taken in viewing these preliminary results for purposes of this discussion. I will only note the variables significant at the .01 level and beyond and then only their sign.

In the independent variable list dummy variables are used in the first and second models for city and in all three models for type of dwelling type. The unspecified condition is Ottawa for city and single detached house for dwelling type.

In 1974 house size, some city variables, total expenditures, age of head and family size and some house types are significant. Large families with high total expenditures living in single detached homes in St. John's consume the most. Western cities consume less than the east, and all other housing types consume less than detached houses, although duplexes not significantly so when number of rooms is controlled for. The unweighted results are similar to the weighted.

When the dependent variable is changed to \$/room and number of rooms is removed from the list of independent variables the general pattern remains. However, family size is no longer significant (probably closely tied to dwelling size only), and education of family head becomes significant with a negative sign. Those with less education consumed more, all other things being equal. Finally duplexes become significant with a positive sign, so when number of rooms is controlled for, duplexes use more energy than detached houses.

In model 3 climatic conditions are taken into account by controlling on degree days in the dependent variable and the list of cities is dropped from the independent variable set.

It should be noted that the value of the coefficients drops so dramatically because there are between 4000 and 7000 degree days in these cities. So the small value of the coefficients does not mean they are unimportant. Total expenditures remains significant as does education of the family head and the rowhouse effect. An important thing to note is the drop in the value of the adjusted R-squared. In fact the independent variables remaining in the equation do not do very much to help in explaining variance in the dependent variable. Other more useful variables should be sought.

When we compare just the unweighted 1974 and 1978 results, in model 1 some change in the Vancouver parameter can be noted and in the importance of semi-detached and duplex housing over detached houses.

In model 2 again the major change is in dwelling type effects. Finally in model 3 only the rowhouse variable shows any difference from the detached:



education of the head is again important, but in 1978 age of head is significant with a positive coefficient. Some improvement is seen in the R-squared for 1978, but it is still very low.

This cross-year comparison from a policy perspective suggests perhaps that improvements have been made in the quality of the detached housing stock in Canada. From a methodological perspective it indicates the importance of choosing the dependent variable with care.

As was earlier noted, much additional analysis and re-analysis will be done using the regression procedures available to refine these results and take the sampling design into account.

As I noted earlier the FAMEX data sets have their limitations but they also contain a wealth of important information which should be fruitfully exploited.

#### REFERENCE

- [1] Mc Dougall, Gordon H.G., Ritchie, J.R. Brent, and Claxton, John D. (1979). "Energy Conservation and Conservation Patterns in Canadian Households: Overview." Behavioral Energy Research Group, 203-2053 Main Hall, University of British Columbia.

Table 1: Energy Consumption Taxonomy - Labels

		Level of In-Home Energy Consumption			
		Low 127 Mil. kJ	Medium 127-222 Mil. kJ	High 222 Mil. kJ	Total
Level of Automobile Gasoline Consump- tion	Low 1136 litre	CHURCH MOUSE 4.5% of sample	9.8% of sample	BEAR 7.5% of sample	11.8
	Medium 1136-4545 litre	14.5% of sample	BEAVER 33.7% of sample	12.3% of sample	60.5
	High 4546 litre	ROADRUNNER 4.0% of sample	12.6% of sample	HIPPO 6.1% of sample	22.6
	Total	23.0	56.1	20.9	100.0

Source: See reference list.



**Table 2: Rank among Typology Cells**

	Churchmice		Hippos		Roadrunners	
	1974	1978	1974	1978	1974	1978
Education of Head (low-hi)	1	1	9	7	7	8
Age (old - yng)	1	2	6	6	9	9
F-T Earners (low-hi)	1	1	8.5	9	7	6.5
Family Size (low-hi)	1	1	9	9	4	4
Child Less than 5 (low-hi)	3	1	4	2	1.5	7
Child 5-15 (low-hi)	1	2.5	7	6.5	5	2.5
Food at Stores (low-hi)	1	1	9	9	4	4
Food at Eating Places (low-hi)	1	1	9	9	6	6
Shelter (low-hi)	1	1	9	7	4	3
Clothing (low-hi)	1	1	9	9	6	5
Personal Care (low-hi)	1	1	9	9	5	4
Medical (low-hi)	1	1	8	8	4	4
Tobacco & Alcohol (low-hi)	1	1	9	9	7	4
Reading, Recreation, Education (low-hi)	1	1	9	8	8	9

**Table 3: Average In-Home Energy Expenditures, 1974-78**

	1974	1978	% Change
Average \$ in-home energy expenditure	451	764	+69
Average \$/room in-home energy expenditure	73	121	+66
Average \$/room/dd in-home energy expenditure	.019	.029	+53

**Table 4: Regression Analysis Results - Model 1 - \$In-Home Energy**

	1974 Unweighted	1974 Weighted	1978 Unweighted
Intercept	197.3 A	225.4 A	298.0 A
No. of Rooms	13.9 A	12.0 A	4.2 C
City - St. John's	193.9 A	204.9 A	341.1 A
Halifax	75.5 A	73.9 B	162.0 A
Montreal	12.2	22.7	-16.6
Toronto	-10.2	-3.0	50.5
Winnipeg	-127.1 A	-125.4 A	-72.2 C
Edmonton	-244.9 A	-243.2 A	-195.8 A
Vancouver	-22.9	-17.5	-71.9 C
Total Expenditures	.006 A	.006 A	.01 A
Age of Head	1.2 A	0.8 B	3.6 A
Family Size	13.2 A	12.1	21.6 B
Education of Head	0.7	0.6	-3.6
House Type - Semi Det.	-50.9 B	-49.0 A	-23.8
Rowhouse	-81.2 A	-88.9 A	-119.7 B
Duplex	-12.3	-13.7	-84.6 C
Adjusted R <sup>2</sup>	0.43	0.34	0.38
F value (prob.)	118.5(.0001)	79.7(.0001)	74.6(.0001)

Note: A = prob. less than .0001, B = prob. less than .001, C = prob. less than .01

Table 5: Regression Analysis Results - Model 2 - \$/Room

	1974 Unweighted	1974 Weighted	1978 Unweighted
Intercept	76.2 A	77.3 A	99.8 A
City - St. John's	30.4 A	32.0 A	74.8 A
Halifax	16.8 A	16.3 B	31.6 A
Montreal	4.5	6.7	6.5
Toronto	-3.5	-1.7	10.1
Winnipeg	-17.6 A	-16.3 A	-0.9
Edmonton	-37.9 A	-36.8 A	-26.4 A
Vancouver	0.3	0.8	-6.7
Total Expenditures	$2.2 \times 10^{-4}$ B	$2.5 \times 10^{-4}$ B	$6.9 \times 10^{-4}$ A
Age of Head	0.015	-0.03	0.33 B
Family Size	0.6	0.04	-0.63
Education of Head	-1.9 A	-1.4 B	-4.0 A
House Type - Semi Det.	-6.5 C	-7.1 B	3.1
Rowhouse	-11.5 A	-11.8 A	-11.0
Duplex	6.1 C	6.6 C	3.24
Adjusted R <sup>2</sup>	.31	.19	.24
F value (Prob.)	73.85(.0001)	38.9(.0001)	41.4(.0001)

Note: A = prob. less than .0001, B = prob. less than .001, C = prob. less than .01.

Table 6: Regression Analysis Results - Model 3 - \$/Room/DD

	1974 Unweighted	1974 Weighted	1978 Unweighted
Intercept	.017 A	.019 A	.02 A
Total Expenditures	$8.01 \times 10^{-8}$ B	$9.4 \times 10^{-8}$ A	$1.4 \times 10^{-7}$ A
Age of Head	$1.8 \times 10^{-5}$	$-7.0 \times 10^{-6}$	$9.9 \times 10^{-5}$ A
Family Size	$-1.4 \times 10^{-5}$	$-18.4 \times 10^{-5}$	$27.0 \times 10^{-5}$
Education of Head	$-5.3 \times 10^{-4}$ A	$-4.7 \times 10^{-4}$ A	$-7.8 \times 10^{-4}$ B
House Type - Semi Det.	$3.4 \times 10^{-4}$	$-7.5 \times 10^{-4}$	$24.8 \times 10^{-4}$
Rowhouse	$-23 \times 10^{-4}$ C	$-35.9 \times 10^{-4}$ A	$-38.8 \times 10^{-4}$ B
Duplex	$16.9 \times 10^{-4}$	$6.3 \times 10^{-4}$	$11.6 \times 10^{-4}$
Adjusted R <sup>2</sup>	.01	.02	.03
F value (Prob.)	5.6(.0001)	6.6(.0001)	9.5(.0001)

Note: A = prob. less than .0001, B = prob. less than .001, C = prob. less than .01

## LOGISTIC REGRESSION ANALYSIS OF LABOUR FORCE SURVEY DATA

S. Kumar and J.N.K. Rao<sup>1</sup>

Standard chisquared ( $X^2$ ) or likelihood ratio ( $G^2$ ) tests for logistic regression analysis, involving a binary response variable, are adjusted to take account of the survey design. The adjustments are based on certain generalized design effects. The adjusted statistics are utilized to analyse some data from the October 1980 Canadian Labour Force Survey (LFS). The Wald statistic, which also takes the survey design into account, is also examined for goodness-of-fit of the model and for testing hypotheses on the parameters of the assumed model. Logistic regression diagnostics to detect any outlying cell proportions in the table and influential points in the factor space are applied to the LFS data, after making necessary adjustments to account for the survey design.

## 1. INTRODUCTION

Logistic regression models have been extensively used by researchers in social, behavioural and health sciences to analyse the variation in binomial proportions (see, for example, the books by Cox (1970) and McCullagh and Nelder (1983)). Due to clustering and stratification used in the survey design the statistical methods for binomial proportions, however, are often inappropriate for analysing sample survey data. For instance, the standard chisquared ( $X^2$ ) or the likelihood ratio ( $G^2$ ) tests greatly inflate the type I error rate (significance level). Hence, some adjustments to the classical methods that take account of the survey design are necessary in order to make valid inferences from survey data. In this article, we have utilized two simple adjustments to  $X^2$  or  $G^2$ , based on certain generalized design effects (deffs) to analyse some data from the October 1980 Canadian Labour Force Survey (LFS) (Section 3). The Wald statistic, which also takes the survey design into account, is also examined.

---

<sup>1</sup> S. Kumar, Census and Household Survey Methods Division, Statistics Canada, and J.N.K. Rao, Department of Mathematics and Statistics, Carleton University.

In addition to formal statistical tests, it is essential to develop diagnostic procedures to detect any outlying cell proportions and influential points in the factor space. Regression diagnostics for the standard linear model have been extensively investigated in the literature (see the recent book by Cook and Weisberg (1982)). Pregibon (1981) recently developed similar methods for the logistic regression with binomial proportions. In Section 4 some of these methods have been applied to the October 1980 LFS data, after making necessary adjustments to account for the survey design.

## 2. THEORETICAL RESULTS

Suppose that the population of interest is partitioned into  $I$  cells (domains) according to the levels of one or more factors, and  $\hat{N}_i$  denotes the survey estimate of the  $i$ -th domain size,  $N_i$  ( $i = 1, 2, \dots, I$ ;  $\sum N_i = N$ ). The corresponding estimate of the  $i$ -th domain total,  $N_{i1}$ , of a binary (0, 1) response variable is denoted by  $\hat{N}_{i1}$ . The ratio estimate,  $\hat{p}_i = \hat{N}_{i1}/\hat{N}_i$ , is used to estimate the population proportion  $\pi_i = N_{i1}/N_i$ .

A logit model on the proportions  $\pi_i$  is given by  $\pi_i = f_i(\beta)$ , where

$$\ln\{f_i/(1 - f_i)\} = \text{logit } f_i = \underline{x}_i' \underline{\beta}, \quad i = 1, \dots, I. \quad (1)$$

In (1),  $\underline{x}_i$  is an  $s$ -vector of known constants derived from the factor levels and  $\underline{\beta}$  is the  $s$ -vector of unknown parameters. Under independent binomial sampling in each domain, the maximum likelihood estimates (m.l.e.) are obtained from the following likelihood equations:

$$X'D(\underline{n}/n)\hat{\underline{f}} = X'D(\underline{n}/n)\hat{\underline{q}}, \quad (2)$$

where  $X' = (\underline{x}_1, \dots, \underline{x}_I)$ ,  $D(\underline{n}/n) = \text{diag}(n_1/n, \dots, n_I/n)$ ,  $\hat{\underline{f}} = \hat{\underline{f}}(\hat{\underline{\beta}}) = (\hat{f}_1, \dots, \hat{f}_I)'$ , and  $\hat{\underline{q}}$  is the vector of sample proportion  $q_i = n_{i1}/n_i$ , where  $n_i$  is the sample size from  $i$ -th domain ( $\sum n_i = n$ ). For general sample designs, we do not have m.l.e. due to difficulties in obtaining appropriate likelihood functions. Hence, it is a common practice to use a "pseudo m.l.e." of  $\underline{\beta}$  or  $\underline{f}$



obtained from (2) by replacing  $n_i/n$  by the estimated domain relative size,  $w_i = \hat{N}_i/\hat{N}$ , and  $\hat{q}_i$  by the survey estimate  $\hat{p}_i$ :

$$X'D(\underline{w})\hat{f} = X'D(\underline{w})\hat{p}. \quad (3)$$

The resulting estimates,  $\hat{\beta}$  and  $\hat{f} = \underline{f}(\hat{\beta})$ , are asymptotically (i.e., in large samples) consistent. The equations (3) may also be written as

$$X'\hat{N}_1(m) = X'\hat{N}_1, \quad (4)$$

where  $\hat{N}_1$  is the vector of estimated counts  $\hat{N}_{i1}$ , and  $\hat{N}_1(m)$  is the vector of pseudo m.l.e.,  $\hat{N}_{i1}(m) = \hat{N}_i \hat{f}_i$ , of the totals  $N_{i1}$ . The estimates  $\hat{\beta}$ , and hence  $\hat{f}$  and  $\hat{N}_1(m)$ , are obtained from (3) or (4) by iterative calculations.

## 2.1 Estimated Variances and Covariances

Let  $\hat{V}$  denote the estimated covariance matrix of  $\hat{p}$ , then the estimated covariance matrix of  $\hat{\beta}$  is given by

$$\hat{D}(\hat{\beta}) = (X'\hat{\Delta}X)^{-1}(X'D(\underline{w})\hat{V}D(\underline{w})X)(X'\hat{\Delta}X)^{-1} \quad (5)$$

in large samples, where  $\hat{\Delta} = \text{diag}(w_1\hat{f}_1(1 - \hat{f}_1), \dots, w_I\hat{f}_I(1 - \hat{f}_I))$ . The diagonal elements of (5) provide the estimated variances of the estimates  $\hat{\beta}_i$ . Similarly, the estimated covariance matrix of the residual vector  $\underline{r} = \hat{p} - \hat{f}$  is given by

$$\hat{D}(\underline{r}) = \hat{A}\hat{V}\hat{A}', \quad (6)$$

where

$$A = I - D(\hat{f})D(\underline{1} - \hat{f})X(X'\hat{\Delta}X)^{-1}X'D(\underline{w}). \quad (7)$$

The diagonal elements  $\hat{V}_{ii}(r)$  of (6) lead to standardized residuals  $r_i/\text{s.e.}(r_i)$  which are useful in detecting outlying cell proportions.

## 2.2 Goodness-of-Fit Tests

The standard chi-squared test of goodness-of-fit of the model (1) is given by

$$\chi^2 = n \sum_{i=1}^I \frac{(\hat{p}_i - \hat{f}_i)^2 w_i}{\hat{f}_i (1 - \hat{f}_i)} = \sum_{i=1}^I \chi_i^2. \quad (8)$$

The likelihood ratio test statistic is given by

$$G^2 = 2n \sum_{i=1}^I w_i \left\{ \hat{p}_i \ln \frac{\hat{p}_i}{\hat{f}_i} + (1 - \hat{p}_i) \ln \frac{(1 - \hat{p}_i)}{(1 - \hat{f}_i)} \right\} = \sum_{i=1}^I G_i^2 \quad (9)$$

Note that  $G_i^2$  is also defined at  $\hat{p}_i = 0$  and 1 as given by  $-2nw_i \ln(1 - \hat{f}_i)$  and  $-2nw_i \ln \hat{f}_i$  respectively. Under independent binomial sampling, it is well known that both  $\chi^2$  and  $G^2$  are asymptotically distributed as a  $\chi^2$  variable with  $I - s$  degrees of freedom, but for general designs this result is no longer valid. In fact,  $\chi^2$  (or  $G^2$ ) is asymptotically distributed as a weighted sum  $\sum \delta_i Z_i$ , of independent  $\chi^2$  variables,  $Z_i$ , each with 1 d.f. where the weights  $\delta_i$  ( $i = 1, \dots, I - s$ ) are the eigenvalues of a "generalized design effects" matrix given by  $\Sigma_0^{-1} \Sigma_\phi$ , where

$$\Sigma_\phi = G'D(\hat{f})^{-1}D(1 - \hat{f})^{-1}VD(\hat{f})^{-1}D(1 - \hat{f})^{-1}G, \quad (10)$$

$$\Sigma_0 = \frac{1}{n} G'\Delta^{-1}G \quad (11)$$

and  $G$  is any  $I \times (I - s)$  matrix of rank  $I - s$  such that  $G'X = 0$ , i.e.,  $G$  is orthogonal to  $X$ . Under binomial sampling,  $\Sigma_0^{-1} \Sigma_\phi$  reduces to  $I$ , the identity matrix

A simple adjustment to  $\chi^2$  (or  $G^2$ ) is obtained (Roberts, 1984) by treating  $\chi_c^2 = \chi^2/\delta$ , or  $G_c^2 = G^2/\delta$ , as  $\chi^2$  with  $I - s$  degrees of freedom (d.f.) under the hypothesis that the model is true, where

$$(I - s)\delta_i = n \sum_{i=1}^I \hat{V}_{ii}(r)w_i / [\hat{f}_i(1 - \hat{f}_i)]. \quad (12)$$

The adjusted statistic  $\chi_C^2$  (or  $G_C^2$ ) should be satisfactory excepting in those cases with a large coefficient of variation (C.V.) of the  $\delta_i$ 's. A better adjustment, based on the Satterthwaite approximation, treats  $\chi_S^2 = \chi_C^2/(1 + a^2)$  or  $G_S^2 = G_C^2/(1 + a^2)$  as  $\chi^2$  with  $(I - s)/(1 + a^2)$  d.f., where

$$a^2 = \sum (\delta_i - \delta_i)^2 / [(I - s)\delta_i^2] \quad (13)$$

is the  $(C.V.)^2$  of the  $\delta_i$ 's and

$$\sum \delta_i^2 = \sum_{i=1}^I \sum_{j=1}^I \hat{V}_{ij}^2(r)(nw_i)(nw_j) / [\hat{f}_i \hat{f}_j (1 - \hat{f}_i)(1 - \hat{f}_j)], \quad (14)$$

where  $\hat{V}_{ij}(r)$  is the  $(i, j)$ -th element of  $\hat{D}(r)$ . The statistics  $\chi_S^2$  and  $G_S^2$  take account of the variation in  $\delta_i$ 's.

A Wald statistic for goodness-of fit of the model (1) is given by

$$\chi_W^2 = \hat{\mathbf{y}}' \mathbf{G} \Sigma_{\hat{\mathbf{y}}}^{-1} \mathbf{G}' \hat{\mathbf{y}}, \quad (15)$$

where  $\hat{\mathbf{y}}$  is the vector of logits  $\hat{y}_i = \text{logit } \hat{p}_i$ . The statistic  $\chi_W^2$  is distributed as  $\chi^2$  with  $I - s$  d.f., in large samples. The statistic  $\chi_W^2$  is not defined if  $\hat{p}_i = 0$  or 1 for some  $i$ . Moreover, it becomes unstable when any  $\hat{p}_i$  is close to 1 (see Section 3), or when the degrees of freedom for  $\hat{\mathbf{V}}$  is not large compared to  $I - s$  (Fay, 1983).

### 2.3 Nested Hypothesis

Suppose the matrix  $X$  is partitioned as  $(X_1, X_2)$  where  $X_1$  is  $I \times r$  and  $X_2$  is  $I \times u$  ( $r + u = s$ ), then the model (1) may be written as

$$\mathbf{y} = X\beta = X_1\beta_1 + X_2\beta_2, \quad (16)$$

where  $\underline{\beta}_1$  is  $r \times 1$  and  $\underline{\beta}_2$  is  $u \times 1$ . We are often interested in testing the null hypothesis  $H: \underline{\beta}_2 = 0$  given the model (16). The "pseudo m.l.e." under  $H$  can be obtained from the equations

$$X_1' D(w) \hat{\underline{f}} = X_1' D(w) \hat{\underline{p}} \quad (17)$$

again by iterative calculations, where  $\hat{\underline{f}} = f(\hat{\underline{\beta}})$ . The standard maximum likelihood ratio tests of  $H: \underline{\beta}_2 = 0$  are given by

$$\chi^2(2|1) = n \sum_{i=1}^I \frac{w_i (\hat{f}_i - \hat{\hat{f}}_i)^2}{\hat{\hat{f}}_i (1 - \hat{\hat{f}}_i)} \quad (18)$$

and

$$G^2(2|1) = 2n \sum_{i=1}^I w_i \left\{ \hat{f}_i \ln \frac{\hat{f}_i}{\hat{\hat{f}}_i} + (1 - \hat{f}_i) \ln \frac{(1 - \hat{f}_i)}{(1 - \hat{\hat{f}}_i)} \right\} \quad (19)$$

respectively. Under binomial sampling, both  $\chi^2(2|1)$  and  $G^2(2|1)$  are asymptotically distributed as  $\chi^2$  with  $u$  d.f. when  $H$  is true, but for general designs this result is no longer valid. In fact  $\chi^2(2|1)$  or  $G^2(2|1)$  is asymptotically distributed as a weighted sum,  $\sum \delta_i(H) Z_i^2$ , of independent  $\chi^2_1$  variates  $Z_i^2$ , where the weights  $\delta_i(H)$  ( $i = 1, \dots, u$ ) are the eigenvalues of the design effects matrix.

$$(\tilde{X}_2' \Delta \tilde{X}_2)^{-1} (\tilde{X}_2' D(w) \hat{V} D(w) \tilde{X}_2), \quad (20)$$

where

$$\tilde{X}_2 = [I - X_1 (X_1' \Delta X_1)^{-1} X_1' \Delta] X_2, \quad (21)$$

(Roberts, 1984). In the binomial case, the design effects matrix  $\tilde{X}_2' \Delta \tilde{X}_2$  reduces to  $I$ , as in the previous case of goodness-of-fit.

A simple adjustment to  $\chi^2(2|1)$  or  $G^2(2|1)$  is obtained by treating  $\chi^2(2|1) = \chi^2(2|1)/\delta_*(H)$  or  $G^2(2|1)/\delta_*(H)$  as  $\chi^2$  with  $u$  d.f. under  $H$ , where

$$u_{\delta}(H) = n \sum_{i=1}^I \tilde{V}_{ii}(r) w_i / \hat{f}_i (1 - \hat{f}_i) \quad (22)$$

and  $\tilde{V}_{ii}(r)$  is the  $i$ -th diagonal element of the covariance matrix of residuals,

$r_i(H) = \hat{f}_i - \hat{f}_i$ , given by

$$\tilde{V}(r) = D(\hat{f})D(1 - \hat{f})\tilde{X}_2\tilde{A}\tilde{X}_2'D(\hat{f})D(1 - \hat{f}) \quad (23)$$

where

$$A = (\tilde{X}_2'\tilde{A}\tilde{X}_2)^{-1}[\tilde{X}_2'D(w)\hat{V}D(w)\tilde{X}_2](\tilde{X}_2'\tilde{A}\tilde{X}_2)^{-1} \quad (24)$$

The standardized residuals  $(\hat{f}_i - \hat{f}_i)/[\tilde{V}_{ii}(r)]^{\frac{1}{2}}$  can also be computed. As in the case of goodness-of-fit, improved approximation, based on Satterthwaite's method, can also be obtained.

A Wald statistic of  $H: \beta_2 = 0$  is given by

$$\chi_W^2(2|1) = \hat{\beta}_2'[\hat{D}(\hat{\beta}_2)]^{-1}\hat{\beta}_2. \quad (25)$$

where  $\hat{D}(\hat{\beta}_2)$  is the principal submatrix in (5) corresponding to  $\hat{\beta}_2$ . Under  $H$ ,  $\chi^2(2|1)$  is asymptotically distributed as  $\chi^2$  with  $u$  d.f. In particular if  $\beta_2$  is a scalar, we can treat  $\hat{\beta}_2/\text{s.e.}(\hat{\beta}_2)$  as  $N(0,1)$ -variate under the hypothesis  $H: \beta_2 = 0$  or  $\hat{\beta}_2^2/\text{var}(\hat{\beta}_2)$  as  $\chi^2$  with 1 d.f.

## 2.4 Diagnostics

It is desirable to make a critical assessment of the logit fit by identifying any outlying cell proportions and influential points in the factor space. For this purpose, the vector of residuals and a projection matrix in the factor space provide useful tools. However, unlike in the case of the standard linear model, the residuals can be defined on different scales. The natural choice that takes account of the survey design is the vector of standardized residuals  $e_i = r_i/[\hat{V}_{ii}(r)]^{\frac{1}{2}}$  given in section 2.1. Since the  $e_i$ 's are

approximately  $N(0, 1)$  under the model (1), the expected numbers of residuals  $e_i$  exceeding 1.96, 2.33 and 2.58 in magnitude are 0.05I, 0.02I and 0.01I respectively, where I is the number of residuals (cells). These expected numbers provide a rough guide to identify any outlying cells. Ignoring the design and hence using standardized residuals under binomial sampling could lead to misleading conclusions.

The standardized residuals  $e_i$ , however, become unreliable for those cells with  $\hat{p}_i = 1$  or close to 1. Following Pregibon (1981), we suggest the use of components of  $X_C^2$  or  $G_C^2$ , viz.,  $\tilde{X}_i = X_i/\delta_i^{1/2}$  or  $\tilde{G}_i = G_i/\delta_i^{1/2}$ ,  $i = 1, \dots, I$ , for residual analysis in order to circumvent this difficulty. In either case, large individual components should roughly indicate cells poorly accounted for by the model. Index plots (i.e., plots of  $\tilde{X}_i$  vs  $i$  and  $\tilde{G}_i$  vs  $i$ ) are useful for displaying these components. Normal probabilities plot of  $\tilde{X}_i$  or  $\tilde{G}_i$  (i.e., the ordered values plotted against standard normal quantiles) is also useful to detect deviations from the model (i.e., deviations from a straight-line configuration).

Pregibon (1981) suggested the use of diagonal elements,  $m_{ii}$ , of the projection matrix

$$\begin{aligned} M &= I - \hat{V}_b^{\frac{1}{2}} X (X' \hat{V}_b X)^{-1} X' \hat{V}_b^{\frac{1}{2}} \\ &= I - H \text{ (say)} \end{aligned} \quad (26)$$

to detect influential points, where  $\hat{V}_b$  is the estimated covariance matrix under binomial sampling, viz.,  $\text{diag}[\hat{p}_1(1 - \hat{p}_1)/(nw_1), \dots, \hat{p}_I(1 - \hat{p}_I)/(nw_I)]$  in the context of survey data. The matrix M arises naturally in solving likelihood equations (4) by iteratively reweighted least squares, and small values of  $m_{ii}$  call attention to extreme points in the factor space. Again, an index plot ( $m_{ii}$  vs  $i$ ) would provide a useful display. It may be noted that the design effect does not come into picture with  $m_{ii}$  since we are using "pseudo m.l.e." based on binomial sampling. Another useful plot which effectively summarizes the information in the index plots  $\tilde{X}_i$  vs  $i$  and  $m_{ii}$  vs  $i$  is given by the scatter plot of  $\tilde{X}_i^2/X_C^2 = X_i^2/X^2$  vs  $h_{ii}$ , where  $h_{ii}$  is the  $i$ -th diagonal element of H given by (26) (see Pregibon, 1981).



The diagnostic measures  $e_i$ ,  $\tilde{X}_i$  or  $\tilde{G}_i$  and  $m_{ii}$  are useful for detecting extreme points, but not for assessing their impact on various aspects of the fit including parameter estimates,  $\hat{\beta}$ , fitted values,  $\hat{f}$ , and goodness-of-fit measures  $X^2/\delta$ , or  $G^2/\delta$ , or others. Following Preqibon (1981) we suggest three measures which quantify the effect of extreme cells (points) on the fit.

(1) Coefficient sensitivity: Let  $\hat{\beta}_j(-\ell)$  denote the pseudo m.l.e. of  $\beta_j$  obtained after deleting the  $\ell$ -th cell data. Then the quantity  $\Delta_j(\ell) = [\hat{\beta}_j - \hat{\beta}_j(-\ell)]/\text{s.e.}(\hat{\beta}_j)$  provides a measure of the  $j$ -th coefficient sensitivity to  $\ell$ -th point. The index plots  $\Delta_j(\ell)$  vs  $\ell$  for each  $j$  provide useful displays but the task of looking at the index plots could become unmanageable if the number of coefficients in the model is large.

(2) Sensitivity of fitted values: Significant changes in coefficient estimates when  $\ell$ -th point (cell) deleted does not necessarily imply that the fitted values  $\hat{f}$  also vary significantly from  $\hat{f}(-\ell)$ , the vector of fitted values obtained after deleting the  $\ell$ -th cell, i.e.,  $\|\hat{f} - \hat{f}(-\ell)\|$  could be small. We therefore use  $[G^2 - \tilde{G}^2(-\ell)]/\delta$ , or  $[X^2 - \tilde{X}^2(-\ell)]/\delta$ , to assess the impact of the  $\ell$ -th point on the fitted values, where  $\tilde{G}^2(-\ell)$  and  $\tilde{X}^2(-\ell)$  are given by (9) and (8) respectively when  $\hat{f}_i = f_i(\hat{\beta})$  is replaced by  $\hat{f}_i(-\ell) = f_i(\beta(-\ell))$ .

(3) Goodness-of-fit: A measure of goodness-of-fit sensitivity is given by  $[G^2 - G^2(-\ell)]/\delta$ , or  $[X^2 - X^2(-\ell)]/\delta$ , where  $G^2(-\ell)$  and  $X^2(-\ell)$  are the likelihood ratio and chi-square statistics obtained after deleting the  $\ell$ -th cell. (Note that  $G^2(-\ell) \neq \tilde{G}^2(-\ell)$ ).

### 3. APPLICATION TO LFS

We have applied the previous methods to some data from the October 1980 Canadian Labour Force Survey (LFS). The sample consisted of males aged 15-64 who were in the labour force and not full-time students. We have chosen two factors, age and education, to explain the variation in unemployment rates via logit models. Age-group levels were formed by dividing the interval [15, 64] into ten groups with the  $j$ -th age group being the interval  $[10 + 5j, 14 + 5j]$ ,  $j = 1, 2, \dots, 10$ , and then using the mid-point of each interval,  $A_j$ , as the value of the age for all persons in that age group. Similarly, the levels of

education.  $E_k$  were formed by assigning to each person a value based on the median years of schooling resulting in the following six levels = 7, 10, 12, 13, 14 and 16. Thus the age by education cross-classification provided a two-way table of  $I = 60$  cell proportions,  $\pi_{jk}$ .

The LFS design employed stratified multi-stage cluster sampling with two stages in the self-representing (SR) urban areas and three or four stages in non-self-representing (NSR) areas in each province. The survey estimates,  $\hat{p}_{jk}$ , were adjusted for post-stratification, using the projected census age-sex distribution at the provincial level. The estimated covariance matrix  $\hat{V}$  of the estimates  $\hat{p}_{jk}$  is based on more than 450 first-stage units (psu's) so that the degrees of freedom for  $\hat{V}$  are large compared to  $I = 60$ .

### 3.1 Formal Tests of Hypotheses.

Scatter plot of the logits  $\hat{v}_{jk}$  vs age levels  $A_j$  at each education level  $E_k$  indicated that  $\hat{v}_{jk}$  for given  $k$  generally increases with age to a maximum and then decreases (i.e., the graph is convex and upward to a maximum). Hence, the following model might be suitable to explain the variation in  $\pi_{jk}$ 's.

$$v_{jk} = \ln \frac{\pi_{jk}}{1 - \pi_{jk}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k + \beta_4 E_k^2$$

$$j = 1, \dots, 10; k = 1, \dots, 6. \quad (27)$$

Some previous work in sociological literature also supports such a model (Bloch and Smith, 1977). Applying the results of Section 2 we obtained the following values for goodness-of-fit statistics

$$\begin{aligned} \chi^2 &= 98.9 & G^2 &= 101.2 \\ \chi^2/\delta_1 &= 52.5 & G^2/\delta_1 &= 53.7, \quad \delta_1 = 1.88. \end{aligned}$$

Since  $\chi^2$  or  $G^2$  is larger than  $\chi_{0.05}^2(55) = 73.3$ , the upper 5% point of  $\chi^2$  with  $I - s = 55$  d.f., we would reject the model if the survey design is ignored. On the other hand, the value of  $\chi^2/\delta_1$  or  $G^2/\delta_1$  indicate that the model is adequate, the significance level (or P-value) being approximately equal

to 0.52. The value of  $\chi^2_S$  when adjusted to refer to  $\chi^2_{0.05}(55)$  is equal to 47.7 which is also not significant. Moreover, in the present context with  $s(= 5)$  relatively small compared to  $I(= 60)$ , the simple correction  $\bar{d}$ , the average cell deff, (see Fellegi, 1980), is very close to  $\bar{\delta}$ :  $\bar{d} = 1.905$  compared to  $\bar{\delta} = 1.88$ : see Rao and Scott (1984) for a theoretical explanation.

The Wald statistic  $\chi^2_W$  is not defined here since two of the cells have  $\hat{p}_{jk} = 1$ , but we made minor perturbations to the estimated counts to ensure that  $\hat{p}_{jk} < 1$  for all cells and then computed  $\chi^2_W$ . The resulting values of  $\chi^2_W$  are all large compared to  $\chi^2/\delta$  (at least 30 times larger than  $\chi^2/\delta$ .) and vary considerably (1715 to 3061). Hence, the Wald statistic is very unstable for goodness-of-fit test in the present context. If the two cells having  $\hat{p}_{jk} = 1$  are deleted, then  $\chi^2_W = 68.4 < \chi^2_{0.05}(53) = 71.0$ , indicating that the model (27) is adequate. However, it is not a good practice to delete cells just to accomodate a chosen test statistic. The other problem with  $\chi^2_W$ , noted by Fay (1983), does not arise here since d.f. for  $\hat{V}$  is large compared to the number of cells in the table.

The pseudo m.l.e., their s.e. and the corresponding s.e. under binomial sampling, all obtained under the model (27), are given in Table 1 along with Wald statistic  $\chi^2_W(2|1)$  and  $G^2$  statistic  $G^2(2|1)/\delta.(H)$  for the hypotheses  $H_1: \beta_i = 0, i=1, 2, 3, 4$  given the model (27). As expected, the true s.e.'s are larger than the corresponding binomial s.e.'s. The hypothesis  $H_4: \beta_4 = 0$  (i.e., coefficient of  $E_i^2$  is zero) is not rejected at the 5% level either by the Wald statistic or  $G^2$  statistic. On the other hand, the coefficient,  $\beta_2$ , of  $A_i^2$  is highly significant. In testing the significance of individual coefficients we compare the values of  $\chi^2_W(2|1)$  or  $G^2(2|1)/\delta.(H)$  to  $\chi^2_{0.05}(1) = 3.84$ , the upper 5% point of  $\chi^2$  - variate with 1 d.f.

We have also tested the following nested hypotheses given model (27):  $H_{34}: \beta_3 = \beta_4 = 0$  (i.e., no education effect);  $H_{24}: \beta_2 = \beta_4 = 0$  (i.e., no quadratic effects). Both  $H_{34}$  and  $H_{24}$  are highly significant:

$$G^2(2|1)/\delta.(H_{34}) = 282.2/1.64 = 172.1, \chi^2_W(2|1) = 165.6 \text{ for } H_{34}:$$

$$G^2(2|1)/\delta.(H_{24}) = 242.2/2.28 = 106.3, \chi^2_W(2|1) = 162.1 \text{ for } H_{24} \text{ compared to } \chi^2_{0.05}(2) = 5.99.$$

Table 1: Pseudo m.l.e.  $\hat{\beta}_i$ , s.e. ( $\hat{\beta}_i$ ),  $\chi^2_W(2|1) = \hat{\beta}_i^2/\text{var}(\hat{\beta}_i)$  and  $G^2(2|1)/\delta_*(H_i)$  Values for the LFS Data under Model (27).

$\hat{\beta}_i$	s.e. ( $\hat{\beta}_i$ )		$\chi^2_W(2 1)$	$G^2(2 1)/\delta_*(H_i)$
	True	Binomial		
0	-2.76	0.557	24.6	
1	0.209	0.0132	250.6	168.4
2	-0.00217	0.000173	157.3	102.1
3	0.0913	0.0891	1.04	1.01
4	0.00276	0.00411	0.45	0.46

Unlike in the case of goodness-of-fit, the Wald statistics is stable for testing nested hypotheses and leads to values close to the corresponding  $G^2(2|1)/\delta_*(H)$  values.

By the above test of goodness-of-fit and tests of nested hypotheses we have arrived at the following simple model involving only four parameters:

$$v_{jk} = \ln \frac{\pi_{jk}}{1 - \pi_{jk}} = \beta_0 + \beta_1 A_{.j} + \beta_2 A_{.j}^2 + \beta_3 E_k, \quad (28)$$

with  $\hat{\beta}_0 = -3.10$ ,  $\hat{\beta}_1 = 0.211$ ,  $\hat{\beta}_2 = -0.00218$  and  $\hat{\beta}_3 = 0.1509$  and corresponding standard errors are 0.247, 0.0130, 0.000172, and 0.0115. We will use the model (28) in Section 3.2 to develop logistic regression diagnostics.

### 3.2 Diagnostics

We now illustrate the use of diagnostics developed in Section 2.4.

#### (i) Residual Analysis

The 60 cells in the two-way table were numbered lexicographically, and the standardized residuals  $e_i$  were computed under the model (28) arrived at above.

formal testing of hypotheses. Among the sixty  $e_i$ , cells numbered 6 and 54 with  $\hat{p}_{jk} = 1$  lead to very large  $e_i$  values: 166.6 and 6.2 respectively. Among the remaining  $e_i$ , the residuals numbers 7, 27 and 59 have values 3.84, 2.73 and 2.52 respectively, whereas the expected number of  $|e_i|$  exceeding 2.33 under model (28) is roughly  $0.02 \times 60 = 1.2$ . Hence, there is some indication that cells 7 and 27 could correspond to outlying cell proportions.

The normal probability plot of  $\tilde{G}_i$  is displayed in FIG. 1: the plot of  $\tilde{X}_i$  is not given to save space since it is similar to the plot of  $\tilde{G}_i$ . Figure 1 indicates no strong deviations from a straight line configuration. The index plot of  $\tilde{G}_i$ , Figure 2, is consistent with Figure 1. Hence, there is no evidence of outlying cell proportions when the components  $\tilde{G}_i$  of  $G_C^2$  are used for residual analysis.

#### (ii) Detection of Influential Cells.

The index plot of  $m_{ii}$  is displayed in Figure 3 which clearly points to cells 1 and 6. Figure 4 displays the plot of  $\tilde{X}_i^2/X_C^2 = X_i^2/X^2$  vs  $h_{ii}$ , where the line with slope - 1 is given by  $X_i^2/X^2 + h_{ii} = 3ave(h_{ii}^*)$ . Here  $h_{ii}^* = h_{ii} + X_i^2/X^2$ , and the values of  $h_{ii}^*$  near unity corresponds to cells which are outlying or influential or both (Pregibon, 1981) and appear above the line in Figure 3. It is clear that cells 1 and 6, and to a lesser extent cells 7 and 58, warrant further examination.

#### (iii) Coefficient Sensitivity.

The index plots for measuring coefficient sensitivity ( $\Delta_j(\ell)$  vs  $\ell$ ) are displayed in Figures 5, 6, 7, and 8 for  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  respectively. It is clear from the plots that cells 2 and 3 cause instability in  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  and  $\hat{\beta}_2$ , whereas  $\hat{\beta}_3$  is affected by cell 7.

#### (iv) Sensitivity of Fitted Values

Figure 9 displays the plot of  $[G^2 - \tilde{G}^2(-\ell)]/\delta_c = c$  vs  $\ell$  for assessing the impact of individual cells on fitted values. Significant peaks in this figure correspond to cells 2 and 3 and to a lesser extent to cell 7. Following Cook (1977) and Pregibon (1981), it may be noted that the comparison of  $c$  to the percentage point of  $\chi^2(s)$  ( $s = 4$  in model (28)) gives a rough guide as to which contour of the confidence region the pseudo m.l.e. is displaced due to deletion of the  $\ell$ -th cell. The value  $c = 2.1$  for cell 2 roughly corresponds to 78% contour of the confidence region.



(v) Goodness-of-fit Sensitivity

Figure 10 displays the plot of  $[G^2 - G^2(-\ell)]/\delta_{\ell}$  vs  $\ell$ : the plot of  $[X^2 - X^2(-\ell)]/\delta_{\ell}$  is similar and hence not displayed but the former plot is preferred (Pregibon, 1981). Significant peaks in this figure corresponds to cells 2, 3, 7, 27, 39 and 54 (values  $\geq 3$ ), the most significant being cell 7 with the value 5.4. By deleting cell 7 and recomputing the adjusted statistic  $G_C^2(-\ell) = G^2(-\ell)/\delta_{\ell}(-\ell)$  where  $\delta_{\ell}(-\ell)$  is the corresponding value of  $\delta_{\ell}$ , we get a value of 48.43 with 55 d.f. compared to  $G^2/\delta_{\ell} = 55.3$  with 56 d.f.

Our investigation on the whole indicated that cells 7, 2 and 3 are possible candidates for deletion, but we feel that their impact is not significant enough to warrant their deletion - one would like to explain the variation among all cell proportions unless certain cells contribute heavily to the disagreement between the data and the fitted model.

ACKNOWLEDGEMENT

We wish to thank M. Gratton of Statistics Canada for producing the graphs included in the paper.

REFERENCES

- [1] Bloch, F.E., and Smith, S.P. (1977). Human capital and labour market employment. J. Human Resources, 12, pp. 550-559.
- [2] Cook, R.D. (1977). Detection of influential observations in linear regression. J. American Statistical Association, 72, pp. 160-170.
- [3] Cook, R.D., and Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, London.
- [4] Cox, D.R. (1970). Analysis of Binary Data. Chapman and Hall, London.



- [5] Fay, R.E. (1983). Replication approaches to the log-linear analysis of data from complex samples. Unpublished manuscript (courtesy of the author).
  
- [6] Felleqi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. J. American Statistical Association, 75, pp. 261-268.
  
- [7] McCullagh, P., and Nelder, J.A. (1983). Generalized Linear Models. Chapman and Hall, London.
  
- [8] Pregibon, D. (1981). Logistics regression diagnostics. Ann. Statist., 9, pp. 705-724.
  
- [9] Rao, J.N.K., and Scott, A.J. (1984). On simple adjustments to chisquared tests with survey data: log-linear and logit models. Unpublished manuscript.
  
- [10] Roberts, G. (1984). On chi-squared tests for logit models with cell proportions estimated from survey data. Unpublished manuscript. Carleton University.

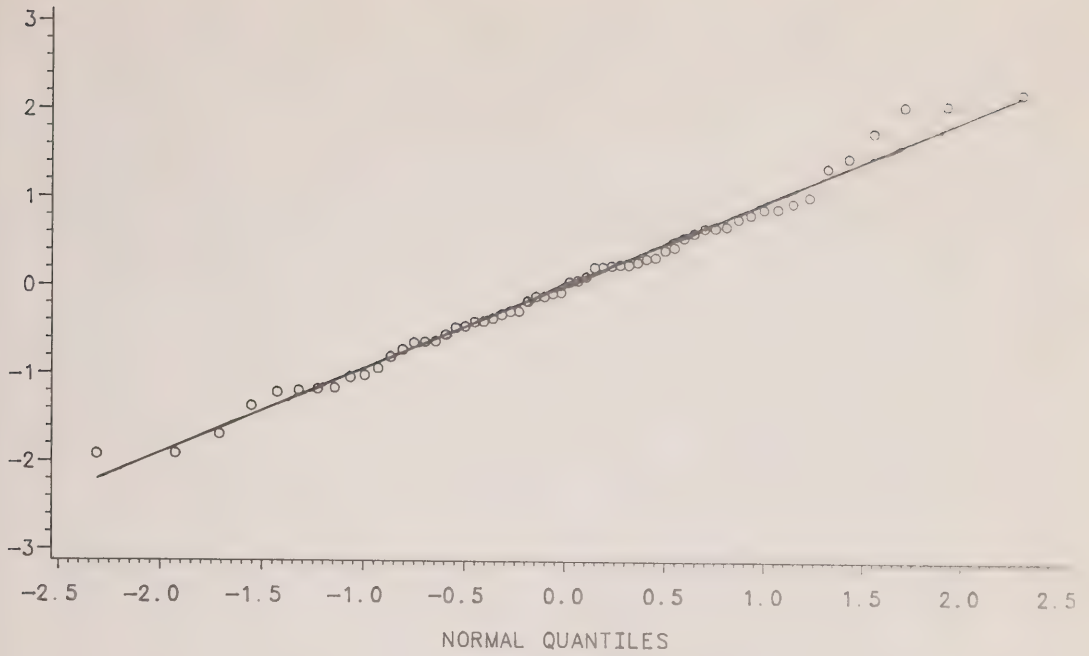


Figure 1: Normal Probability Plot of  $\tilde{G}_i$

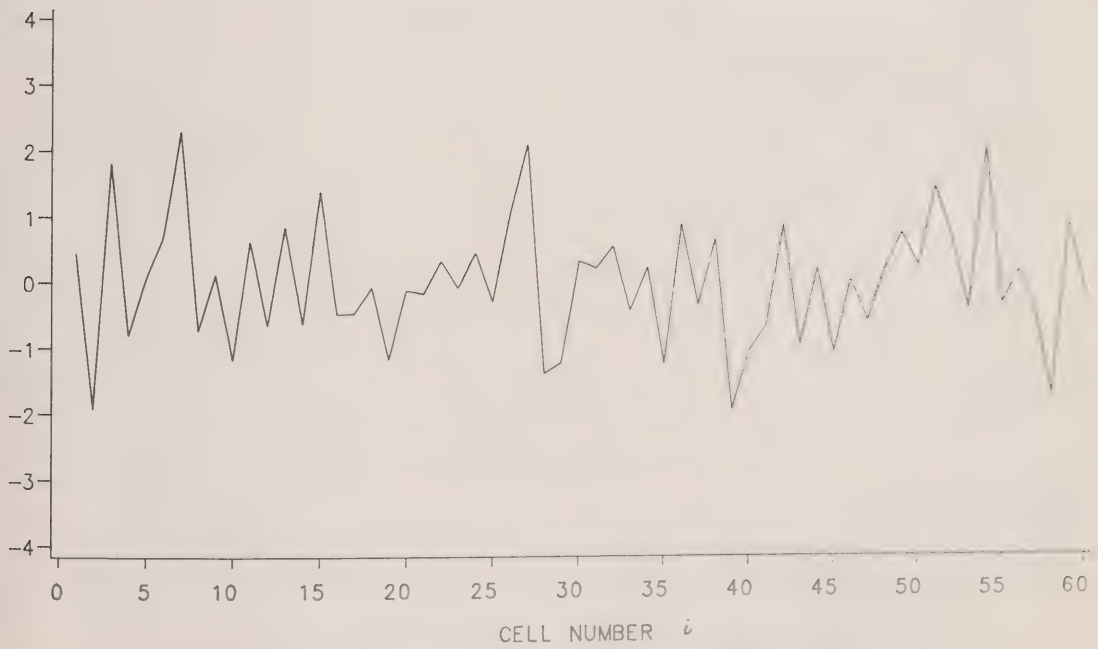


Figure 2: Index Plot of  $\tilde{G}_i$

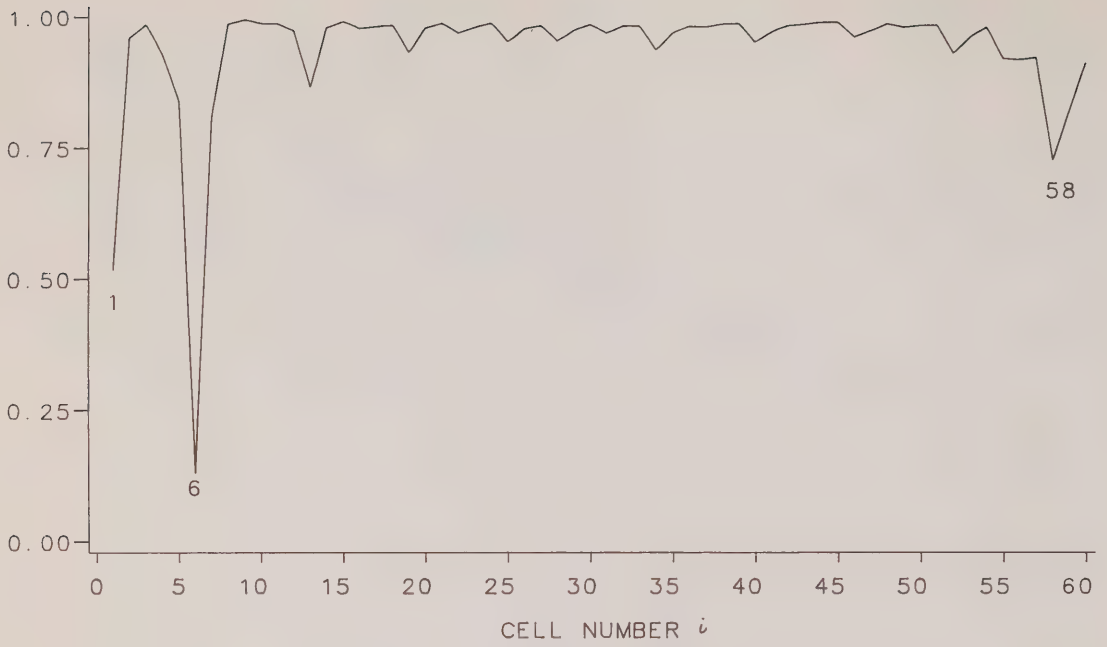


Figure 3: Index Plot of  $m_{ii}$

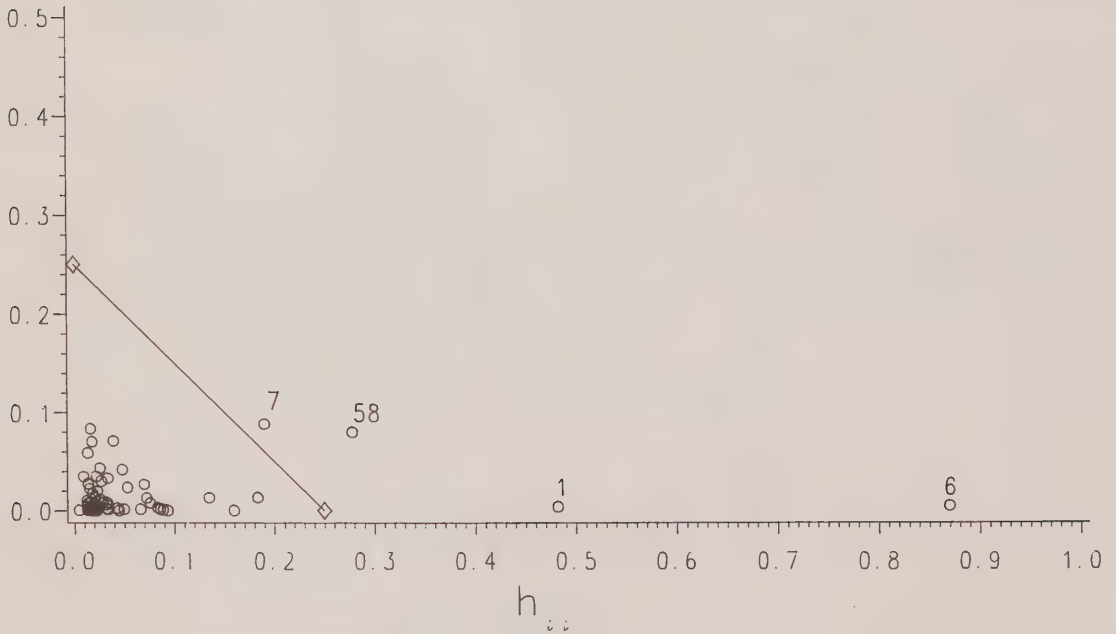


Figure 4: Scatter Plot of  $x_i^2/x^2$  vs.  $h_{ii}$

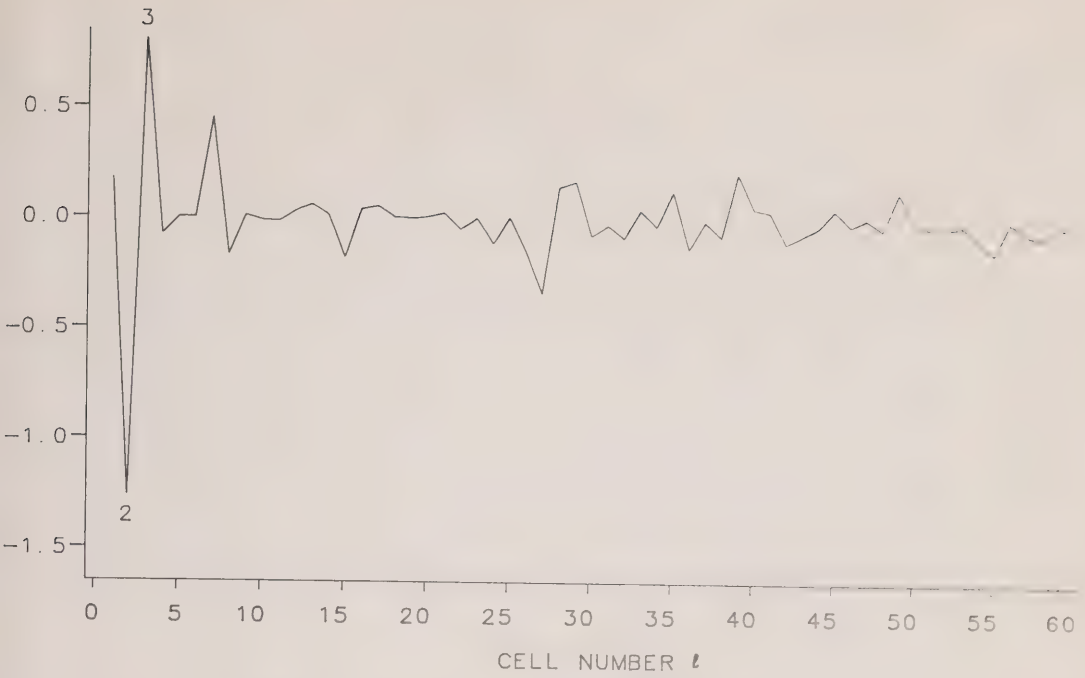


Figure 5: Index Plot of  $\{\hat{\beta}_0 - \hat{\beta}_0(-l)\}/s.e.(\hat{\beta}_0)$

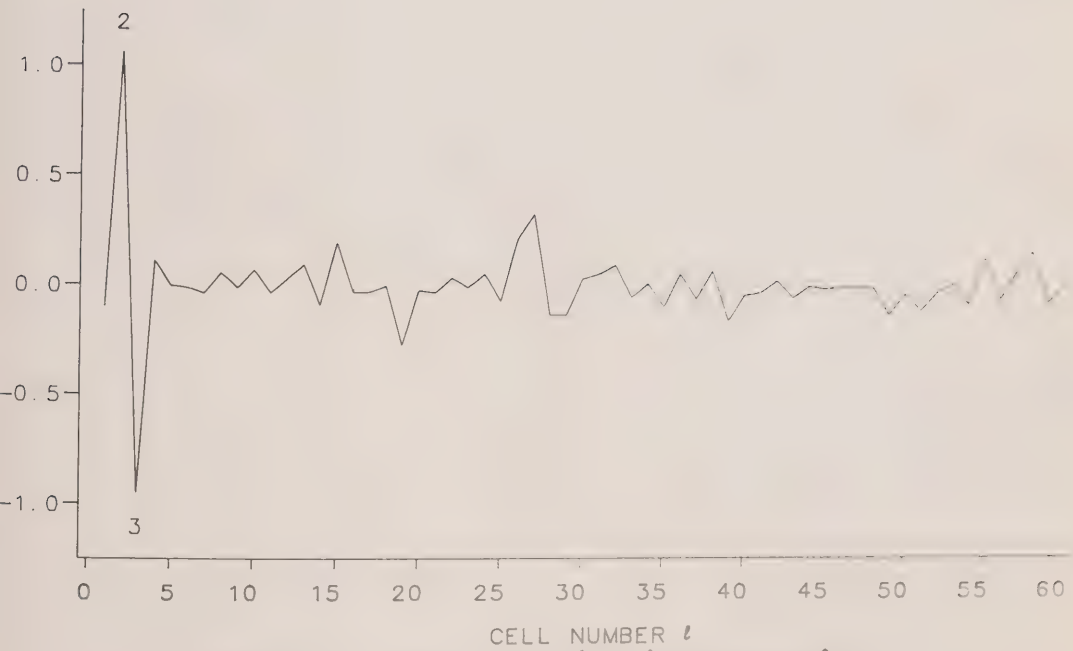


Figure 6: Index Plot of  $\{\hat{\beta}_1 - \hat{\beta}_1(-l)\}/s.e.(\hat{\beta}_1)$

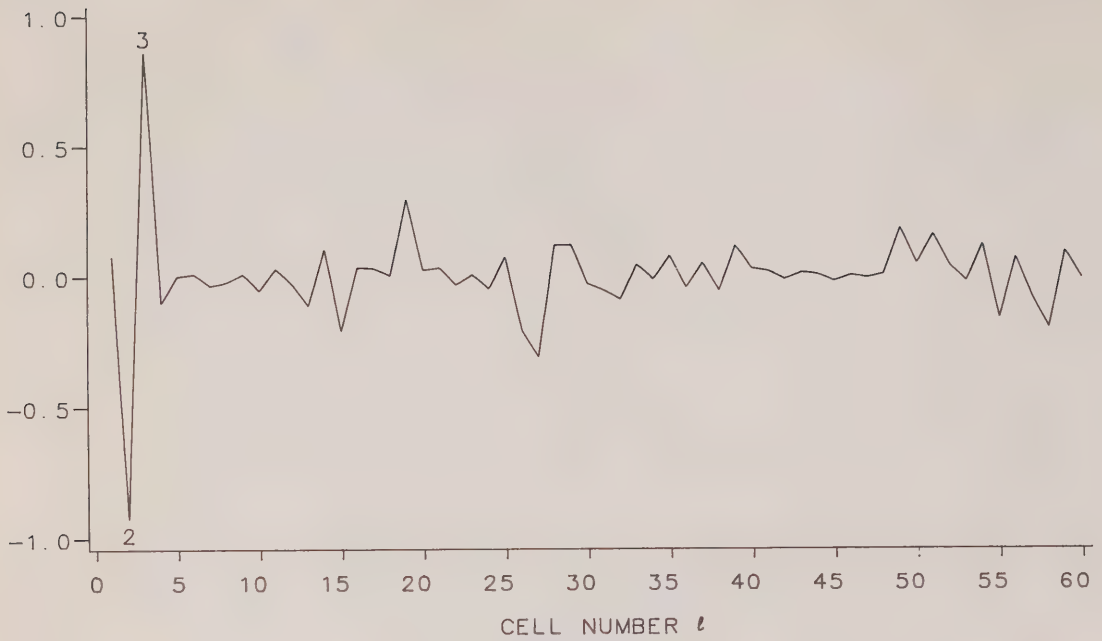


Figure 7: Index Plot of  $\{\hat{\beta}_2 - \hat{\beta}_2(-l)\}/\text{s.e.}(\hat{\beta}_2)$

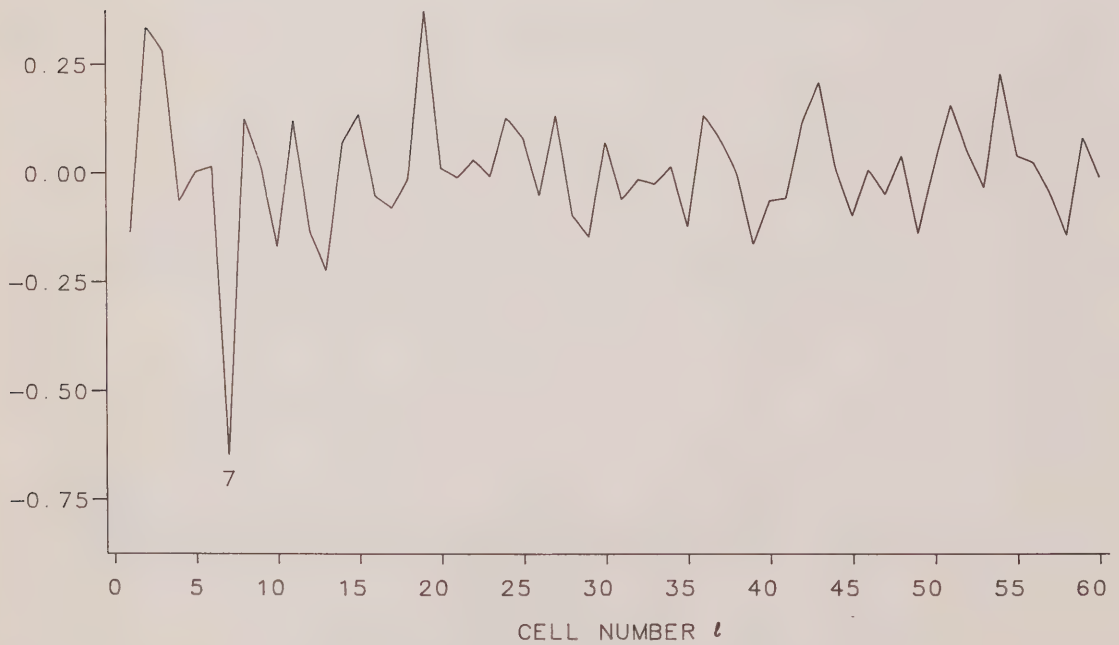


Figure 8: Index Plot of  $\{\hat{\beta}_3 - \hat{\beta}_3(-l)\}/\text{s.e.}(\hat{\beta}_3)$

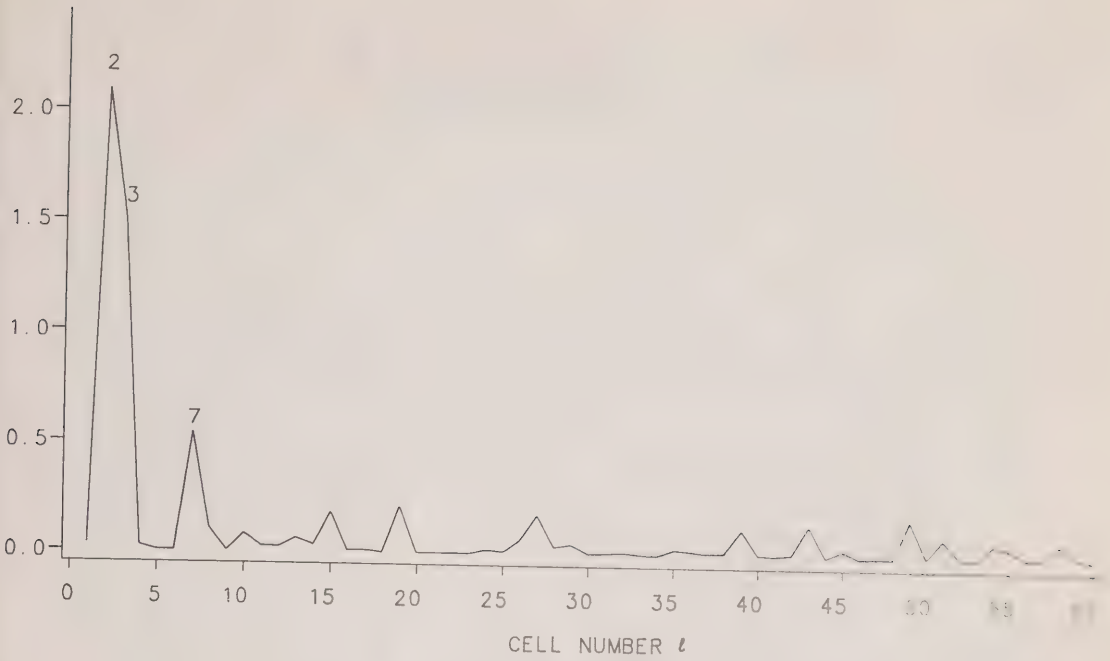


Figure 9: Index Plot of  $\{G^2 - \hat{G}^2(-l)\} / \hat{\delta}$

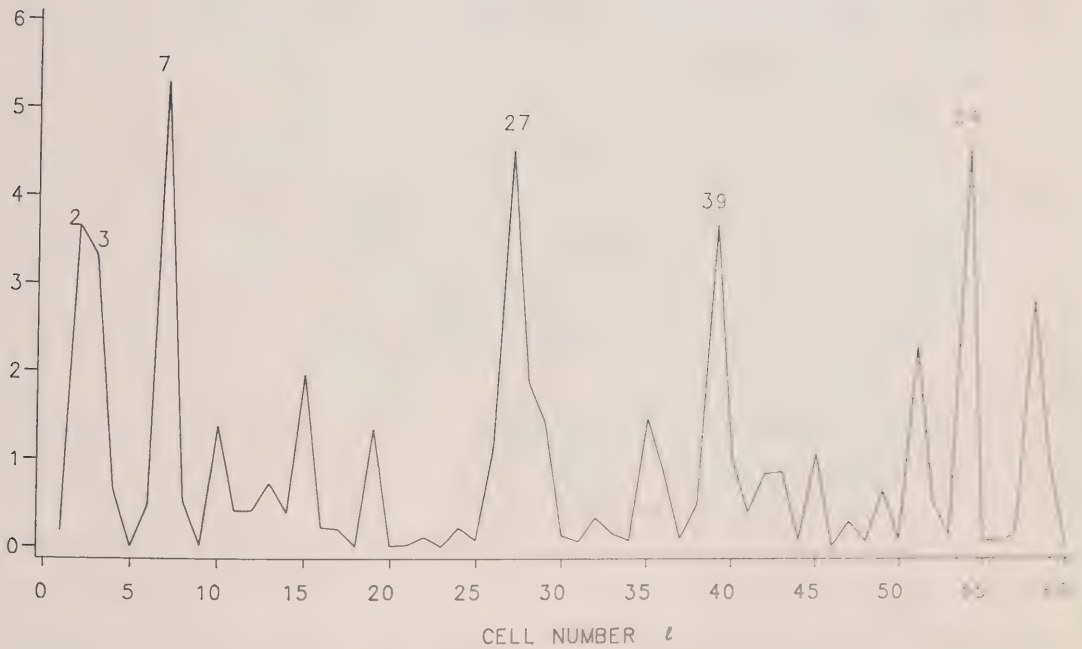


Figure 10: Index Plot of  $\{G^2 - G^2(-l)\} / \hat{\delta}$



## APPLICATION OF LINEAR AND LOG-LINEAR MODELS TO DATA FROM COMPLEX SAMPLES

Robert E. Fay<sup>1</sup>

Most sample surveys conducted by organizations such as Statistics Canada or the U.S. Bureau of the Census employ complex designs. The design-based approach to statistical inference, typically the institutional standard of inference for simple population statistics such as means and totals, may be extended to parameters of analytic models as well. Most of this paper focuses on application of design-based inferences to such models, but rationales are offered for use of model-based alternatives in some instances, by way of explanation for the author's observation that both modes of inference are used in practice at his own institution.

Within the design-based approach to inference, the paper briefly describes experience with linear regression analysis. Recently, variance computations for a number of surveys of the Census Bureau have been implemented through "replicate weighting"; the principal application has been for variances of simple statistics, but this technique also facilitates variance computation for virtually any complex analytic model. Finally, approaches and experience with log-linear models are reported.

### 1. INTRODUCTION

Statistics Canada has played a significant role in many of the methodological developments in the application of analytic methods to sample survey data. The intent of this paper is to review and to share some of the experience acquired by the U.S. Bureau of the Census with these same questions.

The "design-based" (also sometimes called "classical") mode of inference predominates in the analysis and presentation of data by most governmental statistical agencies, such as Statistics Canada and the U.S. Bureau of the Census, as well as by most large private survey organizations. The basis of

---

<sup>1</sup> Robert E. Fay, Statistical Methods Division, U.S. Bureau of the Census, Washington, D.C.

statistical inference with this approach is the randomization employed to select the sample from the finite population. Construction of confidence intervals and tests of hypotheses are based on a large-sample theory tied to this randomization rather than to a specific model. Standard texts such as those by Cochran [4], Kish [17], and Hansen, Hurwitz, and Madow [14] present the elements of this theory. Hansen, Madow and Tepping [15] recently argued the advantages of this approach to the problem of inference from survey data over "model-based" methods; Särndal [25] and Cassel, Särndal, and Wretling [16] have discussed the choice between the model and design-based approaches from a somewhat different point of view. Most of the original development of the design-based theory of inference was specifically for population totals, proportions, means, and ratios, and much of the corresponding literature for the model-based theory similarly concentrates on such basic statistics.

Common analytic models, such as linear regression, log-linear models, and generalized linear models, on the other hand, were initially developed in the context of explicit stochastic models, for example, the normal or multinomial distributions. "Classical" inference here has generally come to refer to statistical inferences based upon such distributional assumptions (where "classical" may include "Bayesian" in this discussion). Developments in "robust" estimation avoid specific distributional requirements, but often maintain assumptions not typically encountered in survey sampling, for example, that the error terms of the model are independent and selected from a symmetric population.

Many researchers familiar with one or more of these analytic models have applied them directly to sample survey data without recognition of the possible consequences of the sample design on the validity of inferences based on the usual distributional assumptions. The subject of this conference, of course, essentially concerns "design-based" alternatives that do reflect the effect of the design. Although all other sections of this paper will address "design-based" methods, the next section considers some of the theoretical and practical issues in choosing between these two approaches, and how these considerations appear manifested in practice at the Census Bureau.

The third section briefly describes some of our experience at the Census Bureau with design-based methods for linear regression. The fourth section discusses an approach taken in the computer implementation of replication

methods, using "replicate weights". Although principally intended for the computation of variance for the usual survey characteristics, this technique also facilitates computation of standard errors for complex models. This general approach may be particularly useful for less standard models, i.e., models other than the linear, log-linear, and other generalized linear models. Finally, some developments with respect to log-linear models are discussed, including specific computer software.

## 2. CHOOSING BETWEEN DESIGN-BASED AND MODEL-BASED INFERENCE FOR ANALYTIC MODELS

The choice between design-based and model-based inference may involve several factors, including effects of stratification, and existence or extent of dependence between sampled values ("clustering"). Many of the essential issues related to this general choice are enumerated by DuMouchel and Duncan [6] in their discussion of whether to incorporate survey weights in linear regression.

If  $\underline{Y}$  represents a column vector of observations  $Y_i$ , and  $\underline{X} = \{X_{ij}\}$ ,  $j = 1, \dots, p$  represents predictors for  $\underline{Y}$ , the model

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon} \quad (2.1)$$

with  $\underline{\varepsilon} = \{\varepsilon_i\}$  composed of independent, identically distributed error terms  $\varepsilon_i \sim N(0, \sigma^2)$ , has as its maximum-likelihood estimate for  $\underline{\beta}$

$$\hat{\underline{\beta}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}. \quad (2.2)$$

Typical survey estimation associates a weight  $W_i$  with each survey case  $i$ , based on the inverse of the probability of selection, often adjusted by factors for nonresponse and ratio estimation. If  $\underline{W}$  represents a diagonal matrix of  $W_i$ , then

$$\hat{\underline{\beta}}_W = (\underline{X}^T \underline{W} \underline{X})^{-1} \underline{X}^T \underline{W} \underline{Y} \quad (2.3)$$

gives a design-consistent alternative incorporating the weights. Under the original stochastic model justifying the choice of (2.2), or, more generally, if the  $\epsilon_i$ 's are uncorrelated with zero expectations and equal variances, (2.3) has a larger sampling variance than (2.2). On the other hand, if these specific assumptions fail (particularly concerning the expectations of the  $\epsilon_i$ 's), (2.3) remains a design-consistent estimate of the census parameter,  $\beta^*$ , defined as the application of (2.2) to the values in the complete finite population, whereas computation of (2.2) for unweighted sample cases cannot guarantee consistent estimation of  $\beta^*$ .

DuMouchel and Duncan further elaborate on the issue of choosing between the variance advantage of (2.2) under the simple model and the consistency of (2.3) under model failure. Their presentation includes a number of citations to earlier commentary by others on both sides of this controversy, and can be recommended for its balanced perspective. Additionally, they propose a test which can be performed with typical computer packages for linear regression, of whether the weighted and unweighted regressions are significantly different. If the test rejects the hypothesis that (2.2) and (2.3) are consistent estimates of the same set of coefficients, then the argument for consistency with the census value,  $\beta^*$ , favors (2.3). If the test does not reject, the authors prefer (2.2) with its (generally) lower variance.

If a researcher rejects (2.2) on the basis of the test proposed by DuMouchel and Duncan, and computes (2.3) instead, the implications of this choice are relatively clear: that (2.3) is selected over (2.2) for its consistency under failure of the model. If the test "accepts" the hypothesis, and (2.2) is used with its associated standard errors derived under the model, caution is nonetheless required in uncritically interpreting (2.2) and associated confidence intervals as statements about the census parameter  $\beta^*$ . In many applications, choice of (2.3) and its associated reliability could be defended as the only "safe" interpretation of the data as an estimate of  $\beta^*$  when model failure is suspected, in spite of possible acceptance by the test of a hypothesis of no significant difference between the weighted and unweighted analyses.

The paper of DuMouchel and Duncan clearly illustrates the most essential consideration in choosing between model-based and design-based inference, namely, efficiency under a correctly specified model versus consistency under



failure of the assumptions of the model. Two footnotes may be added. Although ignoring survey weights is inconsistent under any design-based approach and can only be justified under model-based approaches, not all model-based inference requires ignoring the information represented in the weights.

Rubin [24] gave a concise explanation of this last point in his discussion of the paper of Hansen, Madow, and Tepping [15]. Referring to the more extensive work of Rosenbaum and Rubin [22], Rubin pointed out that a complete Bayesian interpretation of the observed data reflects not only consideration of the functional and distributional relationships in the total population (such as models like (2.1) for the complete population) but also the process by which the sample observations become observed. (In a randomized design, "propensity" to be included in the sample may be equated to probability of selection and the "propensity score" in Rosenbaum and Rubin [22].) On the basis of this consideration, Rubin [23] presented an interesting justification, from a Bayesian perspective, of the use of randomization in sample selection, a procedure that has been staunchly defended by proponents of design-based inference but treated with some disdain by many proponents of model-based inference. Consequently, Rubin advocates model-based inference tempered by careful analysis of the effects of selection or propensity to be included in the sample; these principles in some circumstances could lead to either (2.2) or (2.3), or perhaps alternatives to both.

As a second footnote, DuMouchel and Duncan explicitly restricted their attention to the issue of weighting for stratified simple random sampling. An equally important issue in many applications is the effect on inferences of clustering, that is, dependencies among sampled units due to their joint inclusion in the sample by design, such as persons in sampled households or persons in neighboring households jointly selected into sample. In self-weighting samples (where all sample cases have equal weight), design-based and model-based analyses may often produce the same estimates of the parameters of an analytic model but substantially different assessments of their reliability, unless the dependencies from clustering are explicitly incorporated into the model-based inference. Unlike the issue of the use of weights in stratified simple random samples, where a model-based approach may be defended if the error terms conform to the original full specification of the model, a known dependence among the observations due to clustering (to any serious

degree) inherently conflicts with any assumption of independence of errors that might be required by an overly simplified model. Hence, models that do not reflect known effects of clustering automatically fail to model the data properly.

Design-based inference is the institutional standard at the U.S. Bureau of the Census; yet, practice incorporates both modes of inference with respect to models. Researchers are most likely to adhere strictly to a design-based standard for inferences to national relationships based upon complex samples. When survey weights vary by only a modest degree or not at all, and the effects of clustering may be presumed small, model-based inferences for analytic models appear to enjoy acceptance. The attraction of model-based inference in these cases, no doubt, reflects less a philosophic choice than a practical one: model-based methods are more accessible and familiar than the design-based counterparts. (The author has encountered applications under such conditions on variation on the weights and effects of clustering where design-based methods simply duplicate model-based conclusions, thus justifying the substitution of model-based methods under similar favorable circumstances. When the weights do appreciably vary, or characteristics are subject to considerable clustering, however, examples are easily found where the two modes of inference substantially disagree, and where the model-based inference is highly questionable.)

Specific areas of application at the Census Bureau appear almost exclusively model-based. Methods for imputation of missing data, in particular, some of which derive from explicit parametric models, characteristically avoid any consideration of design-based weights. Another specific field of study, estimation for small areas or domains, often reflects a mixed strategy of design- and model-based inference. Thus, practice at the Census Bureau appears to parallel the choice outlined by DuMouchel and Duncan: efficiency (and simplicity) under the assumed model versus consistency under model failure. Strict inference to national relationships are most likely to elicit design-based methods, while less formal analyses or analyses in which the model is hoped correct (missing data) often favor a model-based approach.



### 3. DESIGN-BASED INFERENCE FOR LINEAR REGRESSION AT THE U.S. CENSUS BUREAU

In general statistical practice, linear regression is probably the single most popular analytic technique. Most data collected by the Census Bureau, particularly for the "demographic areas" involving characteristics of persons or housing, are categorical: linear regression, in any form, is used relatively seldom at the Census Bureau by comparison.

Fuller [13] developed basic results in design-based inference for linear regression, using methods based upon Taylor-series expansions (linearization). These results are incorporated in the computer program SUPER CARP [16], whose development was partially supported by the U.S. Bureau of the Census. We can report successful use of the program ourselves, although it has been applied to only a few problems thus far. The report by Moore [26] is probably the most accessible illustration of the use of SUPER CARP at our institution.

The next section discusses the implementation of replication methods through replicate weights, and we have given preliminary thought, but not yet attempted to implement, alternative computer software specifically designed for this approach. No substantial philosophic difference with SUPER CARP is implied by these considerations, although replication methods tend to give slightly larger and thus more conservative standard errors than linearization. The intent in developing this software would be to take advantage of replication methods developed for some of our surveys, which can be made to reflect the effects of complex estimators more completely than programs implementing linearization.

### 4. COMPUTING DESIGN-BASED VARIANCES THROUGH REPLICATE WEIGHTS

Replication methods, such as jackknife, half-sample, and bootstrap techniques, represent the principal general alternative to linearization for design-based variance estimation for nonlinear statistics. Kish and Frankel [18] presented an early discussion of the use of replication for such purposes and much research has been conducted since.

The popularity of replication for variance estimation has gone through

cycles. Linearization is a powerful technique, of course, and relationships presented by Binder [1] facilitate its implementation for a wide class of analytic models. Census Bureau surveys tend to employ quite complex estimators, however, and fully representing the effect on the sampling variances of these estimators has frequently proven to consume large amounts of professional time, both by statisticians and, especially, experienced computer programmers. Recently, variance computations for a number of surveys have used replication methods achieved through a "replicate weighting" approach. The principal features of this method are to provide a unified approach to enable the computation of variances for a large number of survey characteristics and to simplify the estimation of variance for complex analytic statistics.

The replicate weighting approach is not a new discovery: some of its earlier history is reported in [5], which also describes experience acquired by the U.S. Bureau of Labor Statistics, Bureau of the Census, and Westat, Inc. The algorithm may be said to represent the variance from a (possibly complex) design and a (possibly complex) survey estimator in the form of data to be associated with the survey data file rather than as a set of (possibly complex) variance formulas requiring computer programming. Familiar replication methods, such as balanced half-samples and the jackknife, may be represented through replicate weights, but the algorithm also facilitates the implementation of a much wider class of resampling plans, as in [7]. In [10], it is shown that there exists a resampling plan (actually an infinite number of resampling plans) corresponding to essentially any familiar variance estimator for estimates of population totals, such as variance expressions for multi-stage designs, Yates-Grundy estimators, etc. By representing complex variance relationships as data, variance computation becomes accessible to a larger group of data users.

Estimation in many surveys assigns weights  $W_{i0}$  to each case  $i$ , so that for any characteristic  $X_i$ , estimates of total are given by the weighted sum of the characteristic times the survey weight

$$\hat{X}_0 = \sum_i W_{i0} X_i. \quad (4.1)$$

The product of the replicate weighting approach is a set of additional weights  $W_{ir}$ ,  $r = 1, \dots, R$ , for each survey case  $i$ , from which alternative estimates of total

$$\hat{X}_r = \sum_i W_{ir} X_i \quad (4.2)$$

may be computed. The estimate of variance is given by

$$\hat{\text{Var}}(\hat{X}_0) = \sum_{r=1}^R d_r (\hat{X}_r - \hat{X}_0)^2 \quad (4.3)$$

for predetermined  $d_r$  independent of the choice of survey characteristic  $X$ . (As an example, a simplified balanced half-sample estimate of variance ignoring the effect of any complex survey estimation reflected in the weights  $W_{i0}$ , would be given by assigning weights  $W_{ir}$  equal either to  $2W_{i0}$  or to 0 according to whether case  $i$  was included in half-sample  $r$ , and setting  $d_r = 1/R$  for each  $r$ .) More generally, for a smooth function  $S$  that are functions of weighted population estimates of total  $\hat{X}_0^{(1)}, \dots, \hat{X}_0^{(k)}$ , each of the form (4.1),

$$\hat{\text{Var}}\{S(\hat{X}_0^{(1)}, \dots, \hat{X}_0^{(r)})\} = \sum_{r=1}^R d_r \{S(\hat{X}_r^{(1)}, \dots, \hat{X}_r^{(k)}) - S(\hat{X}_0^{(1)}, \dots, \hat{X}_0^{(k)})\}^2 \quad (4.4)$$

The estimator  $S$  in (4.4) may stand for the sometimes extremely complex estimators often used in survey estimation, incorporating noninterview adjustments and ratio or iterative ratio estimation. Furthermore, these forms of complex survey estimation, if incorporated in the weights  $W_i$ , may be included in the derivation of  $W_{ir}$  as well. Thus, variance computation with this approach falls naturally into three distinct steps or phases:

1. Generate replicate basic weights  $W_{ir}^*$  for the simple unbiased (Horwitz-Thompson) weighting of the data given by the basic weights  $W_{i0}^*$ .
2. Compute replicate (final) weights,  $W_{ir}$ , by applying the same noninter-

view and ratio estimators to the replicate basic weights,  $W_{ir}^*$ , as the original estimation procedures used to compute  $W_{i0}$  from the  $W_{i0}^*$ .

3. Apply (4.4) to the estimation of variance of simple or complex statistics.

The modularity of the preceding three phases is a key feature of this technique: general programs may be used to perform phases 1 and 2, or custom programs may be written to cover unusual circumstances as required. For a single survey, phases 1 and 2 need be performed only once. Programs for phase 3 need take no specific note of the design or estimator and can be run as needed by any user with access to the replicate weights  $W_{ir}$  produced in the second phase.

Although most applications of this method at the Census Bureau have been to estimate variances for basic survey characteristics such as means, totals, or proportions, (4.4) lends itself well to analytic purposes as well. This approach fully represents the effects of complex designs and estimators, whereas in practice implementation of linearization often is restricted to the more common and simple situations. Furthermore, although specific computer software may be developed to implement linearization for common analytic methods, such as linear regression, log-linear models, generalized linear models, etc., formula (4.4) enables researchers to compute variances for more specialized analytic models for which no linearization methods have been programmed, since (4.4) only requires that the researcher apply complete data algorithms to the alternative estimates produced by the replicate weights.

## 5. DESIGN-BASED INFERENCE FOR LOG-LINEAR MODELS

Log-linear models, which express the logarithm of the expected frequencies for categorical responses as a linear function of unknown parameters, encompass both factorial models for cross-classified categorical data, and logistic models for one or more dependent categorical variables as a function of any combination of categorical and continuous predictors. Bishop, Fienberg, and Holland [2] provided one of the earliest books in this rapidly expanding field.

Many log-linear models, particularly those for fully cross-classified categorical data, involve a large number of parameters. The three most typical problems of inference are:

1. To compute standard errors and confidence intervals for the individual estimated parameters,
2. To test the significance of the contribution of specific sets of parameters to the fit of a model,
3. To test the overall goodness-of-fit of the model.

In the context of simple random samples, standard results in maximum likelihood theory provides an answer to these questions, although the Pearson chi-square test rightfully enjoys greater popularity than the likelihood-ratio chi-square test as a solution to the third problem.

Koch, Freeman, and Freeman [19] extended the Weighted Least Squares (WLS) method to complex samples, thereby providing solutions to each of the three principal inferential problems. While this method has proven of substantial general use, it is limited in some applications by the necessity to produce highly precise estimates of the design-based covariance of the sample estimates before the asymptotic theory approximates the actual performance of the WLS procedures. (Further comments on the limitations of WLS are given in [8] and [11].)

Felleqi [12] made an early contribution to the development of alternative tests to WLS for specific situations. More recently, Rao and Scott [20], [21] have formulated and extended a set of related methods to cover the problem of testing for a general class of models including log-linear models. Development of these methods has been closely associated with Statistics Canada.

A less well-known "jackknife chi-square test" [11] gives an alternative approach to the general problem of design-based tests of hypotheses. This test is based upon replication, using (4.4) and a similar expression related to the approximation of the first-order bias (as in the usual jackknife) to draw approximate inferences about the null hypothesis distribution of the usual chi-square tests applied directly to the weighted survey estimates. The method shares much in common with those developed by Rao and Scott. Although a full comparison of the relative merits the jackknifed test and the tests



proposed by Rao and Scott has not been conducted, the preliminary suggestion is that both work well and neither entirely dominates the other. (Further comments are given in [11].)

The jackknifed tests do appear somewhat easier to implement, however, especially to tables involving a large number of cells. A FORTRAN computer program, CPLX (described in [8] and documented by [9]), implementing the jackknifed tests for factorial log-linear models for cross-classified data is now in the public domain. The program also computes replication-based standard errors for parameters of log-linear models, thus also addressing the first of the three problems of inference listed earlier. Although CPLX fits well into an environment in which other survey variances are also estimated through replication approaches, such as the replication weighting techniques described in the previous section, these circumstances are by no means necessary to use the program, and a number of researchers within and outside the Census Bureau have applied the program in a variety of settings.

In time, the author hopes to be able to incorporate the methodology of Rao and Scott into a program like CPLX in order to make both methods available. For the short term, however, the current version of CPLX should be of help to researchers seeking design-based inferences from survey data.

#### REFERENCES

- [1] Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Samples. *International Statistical Review* 51: pp. 279-292.
- [2] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.
- [3] Cassel, C.M., Särndal, C.-E., and Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: John Wiley.
- [4] Cochran, W.G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley.



- [5] Dippo, C.S., Fay, R.E., and Morganstein, D.H. (1984). Computing Variances from Complex Samples with Replicate Weights. Prepared for presentation at the annual meetings of the American Statistical Association, Section on Survey Research Methods.
- [6] DuMouchel, W.H., and Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association 78: pp. 535-543.
- [7] Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- [8] Fay, R.E. (1982). Contingency Tables for Complex Designs: CPLX. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 44-53.
- [9] Fay, R.E. (1983). CPLX - Contingency Tables Analysis for Complex Sample Designs, Program Documentation. Unpublished report, Washington, D.C.: U.S. Bureau of the Census.
- [10] Fay, R.E. (1984). Some Properties of Estimates of Variance based on Replication Methods. Prepared for presentation at the annual meetings of the American Statistical Association, Section on Survey Research Methods.
- [11] Fay, R.E. (1984). A Jackknifed Chi-square Test for Complex Samples. To appear in the Journal of the American Statistical Association.
- [12] Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness-of-Fit Based on Stratified Multistage Samples. Journal of the American Statistical Association 75: pp. 261-268.
- [13] Fuller, W.A. (1975). Regression Analysis for Sample Survey. Sankhyā C 37: pp. 117-132.

- [14] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vols. I and II. New York: John Wiley.
- [15] Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. Journal of the American Statistical Association 78: pp. 776-793.
- [16] Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1978). Super Carp (3rd edition). Ames, 10: Statistical Laboratory, Iowa State University.
- [17] Kish, L. (1965). Survey Sampling. New York: John Wiley.
- [18] Kish, L., and Frankel, M.R. (1974). Inference from Complex Samples. Journal of the Royal Statistical Society, Ser. B 36: pp. 1-37.
- [19] Koch, G.G., Freeman, D.H., and Freeman, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Samples. International Statistical Review 43: pp. 59-78.
- [20] Rao, J.N.K., and Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness-of-Fit and Independence in Two-Way Tables. Journal of the American Statistical Association 76: pp. 221-230.
- [21] Rao, J.N.K., and Scott, A.J. (1984). On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. Annals of Statistics 12: pp. 46-60.
- [22] Rosenbaum, P.R.R., and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies. Biometrika 70: pp. 41-55.
- [23] Rubin, D.B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. Annals of Statistics 6: pp. 34-58.

- [24] Rubin, D.B. (1983). Comment: Probabilities of Selection and Their Role for Bayesian Modeling in Sample Surveys. Journal of the American Statistical Association 78: pp. 803-805.
  
- [25] Särndal, C.-E. (1978). Design-Based and Model-Based Inference in Survey Sampling. Scandinavian Journal of Statistics 5: pp. 27-52.
  
- [26] U.S. Bureau of the Census (1982). Preliminary Evaluation Results Memorandum No. 31: Evaluating the Public Information Campaign for the 1980 Census - Results of the 1980 KAP Survey. Prepared by Jeffrey C. Moore, Washington, D.C.

## LEAST SQUARES AND RELATED ANALYSES FOR COMPLEX SURVEY DESIGNS

Wayne A. Fuller<sup>1</sup>

## 1. INTRODUCTION AND MODEL

Assume that a sample of clusters of elemental units is selected from a finite population divided into  $L$  strata. The total sample of  $n$  clusters (primary sampling units) is given by

$$n = \sum_{h=1}^L n_h \quad (1)$$

where  $n_h \geq 2$  is the number of clusters selected in the  $h$ -th stratum. A column vector of characteristics

$$\tilde{y}_{hij} = (y_{hij1}, y_{hij2}, \dots, y_{hijp})' \quad (2)$$

is observed for the  $j$ -th elemental unit in the  $i$ -th cluster of the  $h$ -th stratum. The vector  $\tilde{y}_{hij}$  is quite general. For example, some elements of the vector can be the powers of products of other entries. Also, one element can be, and often will be, identically equal to one. The cluster totals for the vector are defined by

$$\tilde{y}_{hi.} = \sum_{j=1}^{m_{hi}} \tilde{y}_{hij} \quad (3)$$

where  $m_{hi}$  is the number of elements in the  $hi$ -th cluster.

We shall be interested in the behavior of locally continuous functions of a linear function of the vector of cluster means

---

<sup>1</sup> Wayne A. Fuller, Department of Statistics, Iowa State University.

$$\hat{\theta} = \sum_{h=1}^L W_h n_h^{-1} \sum_{i=1}^{n_h} y_{hi}, \quad (4)$$

where  $W_h$  are fixed weights. Often the weights are

$$W_h = N_h N^{-1}, \quad (5)$$

where  $N_h$  is the number of clusters in the  $h$ -th stratum and  $N$  is the total number of clusters in the population. For the weights (5) the linear function in (4) is the usual unbiased estimator of the finite population mean per cluster. Another set of weights that often is of interest is the set of unit weights

$$W_h = n^{-1} n_h. \quad (6)$$

Our model permits us to consider functions of the mean per element. The usual estimator of the mean per element for a particular  $Y$ -variable is the ratio of the mean per cluster for the  $Y$ -variable to the mean per cluster of the number of elements. The mean number of elements per cluster is the cluster mean of a  $Y$ -variable that is identically one.

Our discussion can be easily expanded to include various forms of subsampling within clusters. Because such expansions add little to the generality of the discussion and add considerable notational complexity, we restrict our attention to single stage sampling within strata.

Our discussion rests heavily on the following central limit theorem for samples from a finite population.

**Theorem 1.** Let  $\{\xi_r: r = 1, 2, \dots\}$  be a sequence of stratified finite populations. Let the population in the  $h$ -th stratum of the  $r$ -th population be a random sample of size  $N_{rh} \geq N_{r-1,h}$  selected from a  $p$  dimensional infinite population with absolute  $2 + \delta$ , where  $\delta > 0$ , moments bounded by  $M_\delta < \infty$ . Let the covariance matrix for the  $rh$ -th infinite population be  $\Sigma_{rh}$ . Let  $L_r \geq L_{r-1}$  be the number of strata in the finite population and let a simple random

sample of  $n_{rh}$  ( $n_{rh} \geq 2$  and  $n_{rh} \geq n_{r-1,h}$ ) units be selected in the  $h$ -th stratum. Let  $f_{rh} = N_{rh}^{-1} n_{rh}$  be a triangular array such that

$$0 \leq f_{rh} < M_{fu} < 1,$$

where  $M_{fu}$  is a fixed number. Let  $\tilde{y}_{rhi.}$  be the total for the  $i$ -th cluster selected in the  $h$ -th stratum for the  $r$ -th population and let

$$\hat{\tilde{\theta}}_r = \sum_{h=1}^{L_r} w_{rh} n_{rh}^{-1} \sum_{i=1}^{n_{rh}} \tilde{y}_{rhi.},$$

$$\tilde{\theta}_{rf} = \sum_{h=1}^{L_r} w_{rh} N_{rh}^{-1} \sum_{i=1}^{N_{rh}} \tilde{y}_{rhi.},$$

$$\tilde{\theta}_r = \sum_{h=1}^{L_r} w_{rh} \mu_{.h..},$$

where  $\tilde{\theta}_{rf}$  is the finite population parameter and  $\mu_{.h..}$  is the mean of the infinite population used to generate the  $h$ -th stratum of the finite population. Assume

$$0 < M_{SL} < \left| n_r \sum_{h=1}^{L_r} w_{rh}^2 n_{rh}^{-1} \tilde{\Sigma}_{rh} \right| < M_{SU} < \infty,$$

where the  $M$ 's are fixed numbers and assume that

$$n_r = \sum_{h=1}^{L_r} n_{rh} \longrightarrow \infty,$$

$$\sup_h \left[ \sum_{t=1}^{L_r} w_{rt}^2 n_{rt}^{-1} \right]^{-1} w_{rh}^2 n_{rh}^{-2} \longrightarrow 0,$$

as  $r \rightarrow \infty$ , where  $w_{rh}$  is a triangular array of weights. Then



$$[\hat{V}\{\hat{\theta}_T - \theta_{TF}\}]^{-\frac{1}{2}}(\hat{\theta}_T - \theta_{TF}) \xrightarrow{L} N(\underline{0}, \underline{I}),$$

$$[\hat{V}\{\hat{\theta}_T - \theta_T\}]^{-\frac{1}{2}}(\hat{\theta}_T - \theta_T) \xrightarrow{L} N(\underline{0}, \underline{I}),$$

where

$$\hat{V}\{\hat{\theta}_T - \theta_{TF}\} = \sum_{h=1}^L w_{rh}^2 (1 - f_{rh}) n_{rh}^{-1} \hat{\Sigma}_{rh},$$

$$\hat{V}\{\hat{\theta}_T - \theta_T\} = \sum_{h=1}^L w_{rh}^2 n_{rh}^{-1} \hat{\Sigma}_{rh},$$

$$\hat{\Sigma}_{rh} = (n_{rh} - 1)^{-1} \sum_{i=1}^{n_{rh}} (\tilde{y}_{rhi.} - \bar{\tilde{y}}_{rh..})(\tilde{y}_{rhi.} - \bar{\tilde{y}}_{rh..})',$$

$$\bar{\tilde{y}}_{rhi.} = n_{rh}^{-1} \sum_{i=1}^{n_{rh}} \tilde{y}_{rhi.}.$$

The proof of this theorem follows from Theorems 1 and 2 of Fuller (1975) and can be extended to multistage samples. Also see Krewski and Rao (1981) and Isaki and Fuller (1982).

Most of our applications are to continuous functions of  $\hat{\theta}_T$ .

**Corollary 1.** Let the assumptions of Theorem 1 hold. Let  $\underline{q}(\underline{\theta})$  be a vector valued function of  $\underline{\theta}$ , where  $\underline{q}(\underline{\theta})$  is continuous with continuous first derivatives for  $\underline{\theta}$  in the sphere  $|\underline{\theta} - \underline{\theta}_T| \leq \delta$  for all  $r$ , where  $\delta > 0$  is fixed. Let  $\underline{G}(\underline{\theta})$  be the nonsingular matrix of first derivatives of  $\underline{q}(\underline{\theta})$ , where the  $ij$ -th element of  $\underline{G}(\underline{\theta})$  is

$$\frac{\partial q_i(\underline{\theta})}{\partial \theta_j},$$

$q_i(\underline{\theta})$  is the  $i$ -th element of  $\underline{q}(\underline{\theta})$  and  $\theta_j$  is the  $j$ -th element of  $\underline{\theta}$ . Then

$$[\hat{G}(\hat{\theta}_T) \hat{V} \{\hat{\theta}_T - \theta_{TF}\} \hat{G}'(\hat{\theta}_T)]^{-\frac{1}{2}} [\hat{g}(\hat{\theta}_T) - \hat{g}(\theta_{TF})] \xrightarrow{L} N(\underline{0}, \underline{I}),$$

$$[\hat{G}(\hat{\theta}_T) \hat{V} \{\hat{\theta}_T - \theta_T\} \hat{G}'(\hat{\theta}_T)]^{-\frac{1}{2}} [\hat{g}(\hat{\theta}_T) - \hat{g}(\theta_T)] \xrightarrow{L} N(\underline{0}, \underline{I}).$$

Corollary 1 is stated for the Taylor estimator of the variance of the approximate distribution of  $\hat{g}(\hat{\theta}_T) - \hat{g}(\theta_T)$ . Suitably defined replication estimators of the variance can also be used. Replication methods include balanced replication methods (see McCarthy (1969)), jackknife methods (See Miller (1974)) and bootstrap methods (see Efron (1979, 1981)). While these methods can be adapted to the sampling situation, the adaptation is not always immediate (see Rao and Wu (1983)).

One class of continuous functions of  $\hat{\theta}$  that deserves special attention is that obtained by using  $\hat{\theta}$  as the dependent variable in a generalized least squares fit.

**Corollary 2.** Let the assumptions of Theorem 1 hold. Let  $\hat{\theta}$  satisfy

$$\hat{\theta} = \hat{h}(\underline{\alpha}).$$

where  $\underline{\alpha}$  is a  $k$ -dimensional vector ( $k \leq p$ ),  $\hat{h}(\underline{\alpha})$  is a continuous function of  $\underline{\alpha}$ , with continuous first and second derivatives for all  $\underline{\alpha}$  in an open sphere containing the true  $\underline{\alpha}_T$  for all  $r$ . Let the parameter space for  $\underline{\alpha}$  be an open bounded subset of  $k$ -dimensional Euclidean space. Let  $\hat{\alpha}_T$  be the vector that minimizes

$$[\hat{\theta}_T - \hat{h}(\hat{\alpha}_T)]' \hat{V}^{-1} \{\hat{\theta}_T - \theta_T\} [\hat{\theta}_T - \hat{h}(\hat{\alpha}_T)].$$

Then

$$[\hat{V} \{\hat{\alpha}_T\}]^{-\frac{1}{2}} (\hat{\alpha}_T - \alpha_T) \xrightarrow{L} N(\underline{0}, \underline{I}),$$

where

$$\hat{V} \{\hat{\alpha}_T\} = [\hat{H}(\hat{\alpha}_T) \hat{V}^{-1} \{\hat{\theta}_T - \theta_T\} \hat{H}'(\hat{\alpha}_T)]^{-1}.$$

and  $\underline{H}(\hat{\underline{\alpha}}_T)$  is the matrix of first derivatives of  $\underline{h}(\underline{\alpha})$  with respect to  $\underline{\alpha}$  evaluated at  $\hat{\underline{\alpha}}_T$ .

## 2. MEANS, RATIOS AND REGRESSIONS

An elementary application of Theorem 1 is the estimation of the mean per cluster and the setting of approximate confidence limits for the mean per cluster. Often the parameter of interest for the mean estimator is the finite population mean per cluster, in which case the finite population correction  $(1 - f_h)$  would be included in the variance estimator.

A slightly more complex application is the estimation of the difference between the means per cluster for two domains. If we let

$$\begin{aligned} Y_{hij1} &= \text{observation on characteristic of interest if element hij is in} \\ &\quad \text{domain 1} \\ &= 0 \text{ otherwise,} \\ Y_{hij2} &= \text{observation on characteristic of interest if element hij is in} \\ &\quad \text{domain 2} \\ &= 0 \text{ otherwise,} \\ Y_{hij3} &= 1 \text{ if element hij is in domain 1} \\ &= 0 \text{ otherwise,} \\ Y_{hij4} &= 1 \text{ if element hij is in domain 2} \\ &= 0 \text{ otherwise.} \end{aligned}$$

the estimated difference between the mean per element in the two domains is

$$\underline{q}(\hat{\theta}) = \underline{q}(\bar{\underline{Y}}_{...}) = \bar{\underline{Y}}_{...3}^{-1} \bar{\underline{Y}}_{...1} - \bar{\underline{Y}}_{...4}^{-1} \bar{\underline{Y}}_{...2}. \quad (7)$$

Two methods of computing the Taylor estimator of variance are often used. The first method computes the estimator of Corollary 1 directly from the matrices  $\underline{G}(\hat{\theta}_T)$  and  $\hat{\underline{V}}\{\hat{\theta}_T - \theta_T\}$  or  $\hat{\underline{V}}\{\hat{\theta}_T - \theta_{TF}\}$ . An algebraically identical computational procedure is to define the observations

$$\hat{z}(y_{hi}, \hat{\theta}) = \hat{z}_{hi} = \hat{g}(\hat{\theta})(y_{hi} - \bar{y}_{h..}) \quad (8)$$

and to compute the ordinary stratified estimator of the variance of the mean per cluster for  $\hat{z}_{hi}$ ,

$$\begin{aligned} \hat{v}\{\hat{z}_{..}\} &= \hat{v}\{g(\bar{y}_{...})\} \\ &= \sum_{h=1}^L w_h^2 (1 - f_h) n_h^{-1} (n_h - 1)^{-1} \sum_{j=1}^{n_h} (\hat{z}_{hi} - \hat{\bar{z}}_{h.}) (\hat{z}_{hi} - \hat{\bar{z}}_{h.}), \end{aligned} \quad (9)$$

where

$$\begin{aligned} \hat{\bar{z}}_{..} &= \sum_{h=1}^L w_h \hat{\bar{z}}_{h.}, \\ \hat{\bar{z}}_{h.} &= n_h^{-1} \sum_{i=1}^{n_h} \hat{z}_{hi}. \end{aligned}$$

For example, the computational form (9) is used in Super Carp. See Hidiroglou et al. (1980, p. 32).

The analyst may be interested in inferences for the particular finite population sampled or for the superpopulation when working with quantities such as differences of means.

One of the more frequent analytic uses of survey data is the computation of regression equations. In fact, the difference between domain means can be expressed as a regression coefficient. Although the vector of regression coefficients is of the form  $\hat{g}(\hat{\theta})$  described in the previous section, it may be advantageous to partition the  $\underline{y}$ -vector of Section 1 into several parts and to give the regression coefficients explicit expressions. The regression equation can be written as

$$y_{hij} = \underline{x}_{hij}' \beta + e_{hij}, \quad (10)$$

where  $y_{hij}$  is the dependent variable, the vector  $\underline{x}_{hij}$  is a  $k$ -dimensional

vector of explanatory variables. The weighted least squares estimator of  $\underline{\beta}$  is

$$\hat{\underline{\beta}}_W = \left[ \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \tilde{x}_{hij} W_{hij} \tilde{x}'_{hij} \right]^{-1} \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \tilde{x}_{hij} W_{hij} y_{hij}. \quad (11)$$

The weights  $W_{hij}$  are permitted to be a function of  $hij$ , but we will assume that the weights are fixed in the sense that they depend only on the elemental identification. This precludes from consideration (except as an approximation) the use of weights that are a function of other elements entering the sample.

Under mild assumptions on the moments of the superpopulation generating the finite population, Theorem 1 is applicable to the estimator defined in (11). If the selection probabilities are denoted by  $\pi_{hij}$ , then the estimator  $\hat{\underline{\beta}}_W$  is a consistent estimator of the finite population vector

$$\underline{\beta}_f = \left[ \sum_{h=1}^L \sum_{i=1}^{N_h} \sum_{j=1}^{m_{hi}} \tilde{x}_{hij} W_{hij} \pi_{hij} \tilde{x}'_{hij} \right]^{-1} \sum_{h=1}^L \sum_{i=1}^{N_h} \sum_{j=1}^{m_{hi}} \tilde{x}_{hij} W_{hij} \pi_{hij} y_{hij}. \quad (12)$$

It follows from (12) that the estimator (11) is a consistent estimator of the finite population regression coefficient when  $W_{hij}$  is proportional to the inverse of the selection probabilities. The error in  $\hat{\underline{\beta}}_W$  as an estimator of  $\underline{\beta}_f$  is

$$\hat{\underline{\beta}}_W - \underline{\beta}_f = \left[ \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \tilde{x}_{hij} W_{hij} \tilde{x}'_{hij} \right]^{-1} \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \tilde{x}_{hij} W_{hij} v_{hij}, \quad (13)$$

where

$$v_{hij} = y_{hij} - \tilde{x}'_{hij} \underline{\beta}_f.$$

By Theorem 1 and Corollary 1 a consistent estimator of the variance of the approximate distribution of  $\hat{\underline{\beta}}_W - \underline{\beta}$  is

$$\hat{V} \{ \hat{\underline{\beta}}_W - \underline{\beta} \} = \hat{A}^{-1} \hat{G} \hat{A}^{-1}. \quad (14)$$

where

$$\hat{\tilde{A}} = \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \tilde{x}_{hij} w_{hij} x'_{hij}.$$

$$\hat{\tilde{G}} = (n - 1)(n - k)^{-1} \sum_{h=1}^L n_h (n_h - 1)^{-1} \sum_{i=1}^{n_h} \hat{\tilde{d}}_{hi} \hat{\tilde{d}}'_{hi}.$$

$$\hat{\tilde{d}}_{hi} = \sum_{j=1}^{m_{hi}} \hat{\tilde{d}}_{hij},$$

$$\hat{\tilde{d}}_{hij} = w_{hij} \tilde{x}_{hij} \hat{v}_{hij},$$

$$n = \sum_{h=1}^L \sum_{i=1}^{n_h} m_{hi}.$$

$$\hat{v}_{hij} = y_{hij} - \tilde{x}_{hij} \hat{\beta}_w.$$

and  $\hat{\beta}$  is the superpopulation analog of  $\beta_f$ . This particular form of the estimator of variance was suggested by Fuller (1975) and is used in Super Carp.

One of the frequently asked questions faced by survey statisticians is: "In computing the regression equation, should I use the sampling weights?" As with most such questions, the answer is "It depends." The fact that the question is asked generally means that the questioner has in mind inference for a population beyond the finite population sampled. This does not mean that the particular superpopulation is completely defined or definable. It does suggest that the questioner is postulating that the finite population is generated by a superpopulation in which some type of linear model holds. One quantification of the hypothesis that weights are not required is the superpopulation hypothesis

$$H_0: \theta_\pi = \theta_{(1)}. \quad (15)$$

where the  $\theta$ 's are superpopulation analogs of (12),

$$\theta_\pi = \left[ \sum_{h=1}^L \sum_{i=1}^{N_h} E_\xi \sum_{j=1}^{m_{hi}} \{x_{hij} \pi_{hij} x'_{hij}\} \right]^{-1} \sum_{h=1}^L \sum_{i=1}^{N_h} E_\xi \left\{ \sum_{j=1}^{m_{hi}} \tilde{x}_{hij} \pi_{hij} y_{hij} \right\},$$



$$\theta_{(1)} = \left[ \sum_{h=1}^L \sum_{i=1}^{N_h} E_{\xi} \left\{ \sum_{j=1}^{m_{hi}} \tilde{x}_{hij} \tilde{x}'_{hij} \right\} \right]^{-1} \sum_{h=1}^L \sum_{i=1}^{N_h} E_{\xi} \left\{ \sum_{j=1}^{m_{hi}} \tilde{x}_{hij} y_{hij} \right\}, \quad (16)$$

and  $E_{\xi}$  denotes expectation with respect to the superpopulation. This is a testable hypothesis. It seems that, at a minimum, a test of this hypothesis should be constructed if one performs an unweighted analysis of a sample with unequal selection probabilities.

If the null hypothesis also includes the hypothesis that the estimator with unit weights is the minimum variance estimator, then the test of the hypothesis is given by the statistic

$$F_{n-L-2k}^k = k^{-1} \hat{\delta}_2' \hat{V}^{-1} \hat{\delta}_2. \quad (17)$$

where

$$(\hat{\delta}_1', \hat{\delta}_2')' = \left[ \sum_{h=1}^L \sum_{i=1}^{N_h} \sum_{j=1}^{m_{hi}} \tilde{z}_{hij} \tilde{z}'_{hij} \right]^{-1} \sum_{h=1}^L \sum_{i=1}^{N_h} \sum_{j=1}^{m_{hi}} \tilde{z}_{hij} y_{hij},$$

$$\tilde{z}_{hij} = (\tilde{x}'_{hij}, \tilde{x}_{hij} w_{hij}),$$

and

$$\hat{V} = \begin{pmatrix} \hat{V}_{11} & \hat{V}_{12} \\ \hat{V}_{21} & \hat{V}_{22} \end{pmatrix} \quad (18)$$

is defined by (14) with  $\tilde{z}_{hij}$  replacing  $\tilde{x}_{hij}$ . As the notation suggests, the statistic is approximately distributed as Snedecor's F with k and  $n - L - 2k$  degrees of freedom.

**Example 1.** Table 1 contains observations on 37 area segments collected by the Statistical Reporting Service, U.S. Department of Agriculture in northcentral Iowa in 1978. Two determinations on the hectares of soybeans are reported. The first is obtained by personal interview in the June Enumerative Survey. The second is obtained from a classification of Landsat data based upon a classifier developed by the Statistical Reporting Service. The original objective of the study was to use the Landsat data to construct a regression

estimator of the total acres. We use the data to illustrate the computation of regression statistics from survey data. The sample most nearly approximates a stratified sample with strata identified in the column headed "county". The inverse of the sampling rates is given in the weight column. The estimated regression equation for the regression of interview hectares on satellite hectares defined by estimator (11) is

$$\hat{Y} = -11.845 + 1.1602X,$$

(8.332) (0.0922)

where the numbers in parentheses are the standard errors obtained from the estimated covariance matrix calculated by equation (14).

Calculations were performed using Super Carp. If the equation and standard errors are calculated using unit weights in equations (11) and (14), respectively, we have

$$\hat{Y} = -3.927 + 1.0850X.$$

(9.282) (0.0963)

If we calculate the F-test suggested in equation (17), we obtain

$$F_{23}^2 = 2.81.$$

At first glance, this test is large enough to cause to suspicion about the equality of the two coefficients. Because this sample is very small and because of the structure of the weights, the test is nearly a test between two lines, the line for county one, and the average line for the remaining counties. In this small sample the deviations from the line in county one are small. Hence, the estimated standard errors of the coefficients for the two added variables are small. This phenomenon is discussed further in Section 3. If one uses the ordinary regression F-test that assumes homogeneous error variances and ignores the stratification, one obtains

$$F_{33}^2 = 0.68.$$

While this statistic is not distributed as Snedecor's F, it does make one feel more comfortable with the assumption that the two weighting procedures are estimating the same equation.

Table 2 contains the standard errors of regression coefficients estimated under alternative assumptions. The estimated standard errors for the intercept behave much as one might anticipate. The stratified weighted sample procedure has the smallest estimated standard error followed by the stratified unit weight procedure and the ordinary least squares procedure. Do not forget these are estimated standard errors. The two stratified procedures are consistent under the stratified model. The weighted estimator has smaller variance because the observations for stratum 1, the stratum with the largest weight, lie closer to the estimated line than do the points in other strata. The ordinary least squares estimated standard error is not consistent under the stratified model. If the sample is treated as a cluster sample of counties, the estimated standard errors for the intercept are about 30 to 40 percent larger than the corresponding values for the stratified sample.

The estimated standard errors for the slope display a different behavior. The smallest estimated standard error is associated with the unit weight cluster estimation, and the largest estimated standard error is associated with ordinary least squares. Roughly speaking, the variation of slopes among clusters is small relative to the within cluster variation. Because the weights are inversely correlated with the observed variability, the weighted estimators have smaller estimated variances. This is a small sample, but it is sufficient to demonstrate that unit weights do not always produce smaller variances than sample weights and that stratification and clustering can have rather complex effects on the estimated variances of the regression coefficients.

### 3. WHAT IS A LARGE SAMPLE?

Our discussion has rested on the large sample properties of estimators and of estimators of variance. If the limiting normal distribution is being used to establish confidence intervals, the size of the sample required for a good approximation depends upon the nature of the original population. For

example, if the characteristic is a rare zero-one item (probability less than 0.05, say), a very large sample (more than 1,400 for a simple random sample (Cochran, 1977, p. 58)) will be required for the normal approximation. The binomial with small  $p$  is only one example of the very skewed populations often encountered in sampling practice. Measures of size such as gross sales of firms, number of employees of firms, number of animals per farm, and family income are examples of skewed populations for which large samples are required before the distribution of the mean approaches normality. On the other hand, the distribution of the mean for items such as family size may approximate the normal distribution for small (less than 100) sample sizes.

The use of the Taylor expansion is semi-nonparametric in that the approximation holds, in large samples, under very mild assumptions on the population. The large sample requirements are met if we have no isolated points in our sample space. The method may perform poorly in situations where the generating distribution and sample size are such that an observation or observations are isolated from the remaining cluster of points. We consider the problem of estimating the variance of the vector of regression coefficients used to test the effect of weighting on the coefficients in the soybean example. The original vector is

$$(1, X, XW, W).$$

and the hypothesis to be tested is the hypothesis that the coefficients for  $XW$  and  $W$  are zero. To illustrate the problems associated with variance estimation for the vector of coefficients for the soybean data set, we create a vector that is orthogonal in the unit weight metric. The matrix of observations on the transformed independent variables is composed of the residuals obtained in the regression of each variable, except the first, on the elements preceding it in the original vector. Table 3 contains the transformed regression variables  $(X - \bar{X}, RWX, RW)$ . Only a few digits have been retained to make it easier to read the table.

When we regress  $Y$  on  $(1, X - \bar{X}, RWX, RW)$  we obtain

$$\hat{Y} = 95.34 + 1.085(X - \bar{X}) + 0.093 \times 10^{-2}RWX - 0.015RW,$$

(2.24)	(0.093)	(0.044)	(0.023)
--------	---------	---------	---------

where the estimated standard errors were computed for a stratified sample with unit weights using expression (14). If the regression and standard errors are computed by ordinary least squares, we obtain

$$\hat{Y} = 95.34 + 1.085(X - \bar{X}) + 0.093 \times 10^{-2}RWX - 0.015RW.$$

(3.37) (0.113)                      (0.086)                      (0.034)

The estimated standard error for the coefficient of RWX obtained by Taylor methods is about one half of that obtained by ordinary least squares methods. This can be explained by the data configuration.

The first observation on RWX is much larger in absolute value than any other observation. Of the total sum of squares for RWX, 67 percent is due to this observation. The Taylor approximation to the variance uses the sample variance of deviates called  $\hat{d}_{hij}$  in (14) to estimate the variance of the statistic. The deviations from regression, denoted by  $\hat{v}$ , are given in the last column of Table 3. The  $\hat{v}$  value for observation one is among the smaller values. The mean square for the residuals is 421. The product  $(RWX)(\hat{v})$  for the first observation is -1113. This product is of the same order of magnitude as the product for observations 3, 33 and 36. Therefore, while the first observation is responsible for about 67 percent of the sum of squares of RWX, it is responsible for only about 15 percent of the sum of squares of  $(RWX)(\hat{v})$ . This is because  $\hat{v}^2$  for the first observation is less than one tenth of the average of the squares of the other observations. Furthermore, the squared deviation for the first observation is biased downward because the method of least squares will cause the estimated plane to pass close to an observation that is separated from the other observations. Thus, if all of the observations have the same error variance, the Taylor method will produce an estimate of the variance of the coefficient for RWX that is biased downward.

Did the procedure underestimate the variance for this sample? We do not know. If we use the parametric procedure of ordinary least squares, we assign the pooled estimate of error variance to the separated observation. It is not possible to determine if this procedure is correct because our estimate of variance for the separated observation is a one degree of freedom estimator.



In this situation most people will feel more comfortable assuming that the variance for the separated point is the same as the variance of the other points rather than taking the small observed variance of the single point.

In the nonparametric world a single observation contains little information about the variability of the population that generated the observation. Furthermore, an observation separated from other observations is essentially a single observation. In the full parametric world the separated observation is in the fold because the separated observation is specified to have been created by the same generating mechanism that created the other observations. For data of the type displayed in Table 3, the answer obtained by parametric methods rests very heavily on assumptions about the error variance.

In the estimation of variances, one measure of the numerical size of the sample is the number of cluster degrees of freedom. Thus, for example, the estimated covariance matrix for a k-dimensional vector random variable is singular unless

$$\sum_{h=1}^L (n_h - 1) > k.$$

In setting approximate confidence intervals it seems reasonable to use Student's t distribution with degrees of freedom no greater than  $\sum (n_h - 1)$ . Because the variance of an estimated variance is a function of the fourth moments of the population, estimated variances are notoriously unreliable. The coefficient of variation for the squares is  $2^{\frac{1}{2}}$  for the normal and considerably larger for many other common distributions.

If the error variances in the strata are unequal or if unequal weights are applied to the estimates of different strata, the variance of the variance estimator can be considerably different from that suggested by a simple calculation of error degrees of freedom. Table 4 has been constructed using the data configurations of Table 1 to illustrate these effects on the estimated variance. In the first column we assume that stratification is ineffective in that we assume each stratum variance is equal to the variance of the population. We assume the parent population to be normal so that we can give an explicit expression for the variance of the variance. In this situation stratification produces an estimated error variance for a mean with a variance



that is proportional to  $(26.6)^{-1}$  while a simple random sample produces a variance of the estimated variance that is proportional to  $36^{-1}$ . The effective degrees of freedom for the stratified sample is slightly less than 27 because of the unequal sample sizes within strata. If we use the sample weights of Table 1 and the usual stratified variance estimator, the variance of the estimated variance is proportional to  $(4.6)^{-1}$ . This large reduction is due to the large weight for the first stratum. If the variance in the first stratum is one half of the variance in other strata, then the effective degrees of freedom for the variance estimator is 12.4. In the last column we give the effective degrees of freedom for the simple random sample if the variance of the simple random sample is twice that of the stratified sample. This illustrates the fact that stratification can reduce both the variance of the estimated mean and the variance of the estimated variance of the mean.

While we are unable to specify the number of error degrees of freedom required for our approximations, it is clear that we shall be uncomfortable with a small number of degrees of freedom, particularly with unequal weights.

The theory of Corollary 1 uses a linear approximation to the nonlinear function of the sample means to approximate the behavior of the nonlinear function. If this approximation is to perform well, the curvature of the function must be small relative to the standard error of the sample means. For example, if the function is quadratic

$$q(\bar{Y}) = \alpha_1 \bar{Y} + \alpha_2 \bar{Y}^2,$$

the linear approximation is

$$g(\bar{Y}) \doteq \alpha_1 \mu + \alpha_2 \mu^2 + (\alpha_1 + 2\alpha_2 \mu)(\bar{Y} - \mu).$$

The expected value of  $g(\bar{Y})$  is

$$E\{g(\bar{Y})\} = \alpha_1 \mu + \alpha_2 [\mu^2 + V\{\bar{Y}\}].$$

For the linear approximation to perform well we must have small  $V\{\bar{Y}\}$  and/or

small  $\alpha_2$ .

In summary, to be comfortable with the use of large sample theory we require:

1. A reasonable number of observations in the sense that no observations are widely separated from the main clusters of observations. This is another way of saying that the Taylor deviates are such that the mean of the deviates is nearly normally distributed.
2. A reasonable number of effective error degrees of freedom for the estimator of variance.
3. The curvature of the nonlinear function of sample means to be small relative to the standard error of the sample means.

#### ACKNOWLEDGEMENTS

This research was partly supported by Research Agreement 58-319T-1-0054X with the Statistical Reporting Service of the U.S. Department of Agriculture. I thank Nancy Hasabelnaby for computations and Carol Francisco for comments.

#### REFERENCES

- [1] Cochran, W.G. (1977). Sampling Techniques 3rd Ed. Wiley, New York.
- [2] Efron, B. (1979). Bootstrap method: Another look at the jackknife. Ann. Statist. 7, pp. 1-26.
- [3] Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika 68, pp. 589-599.
- [4] Fuller, W.A. (1975). Regression analysis for sample survey. Sankhyā Series C 37, pp. 117-132.
- [5] Fuller, W.A. and Hidiroqlou, M.A. (1978). Regression estimation after correcting for attenuation. J. Amer. Statist. Assoc. 73, pp. 99-104.

- [6] Hidiroqlou, M.A., Fuller, W.A., and Hickman, R.D. (1980). Super Carp, Department of Statistics, Iowa State University, Ames, Iowa.
- [7] Isaki, C. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. J. Amer. Statist. Assoc. 77, pp. 89-96.
- [8] Kish, L. and Frankel, M.R. (1974). Inference from complex samples. J. Roy. Statist. Soc. B 36, pp. 1-22.
- [9] Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. Ann. Statist. 9, pp. 1010-1019.
- [10] McCarthy, P.J. (1965). Stratified sampling and distribution-free confidence intervals for the median. J. Amer. Statist. Assoc. 60, pp. 772-783.
- [11] McCarthy, P.J. (1969). Pseudo-replication: Half-samples. Rev. Int. Statist. Inst. 37, pp. 239-264.
- [12] Miller, R.G., Jr. (1974). The jackknife - a review. Biometrika 61, pp. 1-15.
- [13] Rao, J.N.K. and Wu, C.F.J. (1984). Bootstrap with stratified samples. Technical Report No. 19 of the Laboratory for Research in Statistics and Probability. Carleton University, Ottawa, Canada.

**Table 1: Soybean Area Determined by Two Methods**

County	Segment	Weight	Soybean Hectares	
			Interview (Y)	Satellite (X)
1	1	502	8.09	24.75
1	2		106.03	98.10
1	3		103.60	112.50
2	1	212	6.47	43.20
2	2		63.82	80.10
3	1	188	43.50	61.65
3	2		71.43	92.70
3	3		42.49	74.25
4	1	190	105.26	98.10
4	2		76.49	99.45
4	3		174.34	152.10
5	1	134	95.67	57.60
5	2		76.57	66.15
5	3		93.48	91.80
6	1	189	37.84	34.65
6	2		131.12	97.65
6	3		124.44	116.10
7	1	172	144.15	136.35
7	2		103.60	99.45
7	3		88.59	99.90
7	4		115.58	123.30
8	1	114	99.15	85.50
8	2		124.56	121.50
8	3		110.88	77.40
8	4		109.14	102.60
8	5		143.66	133.65
9	1	193	91.05	75.15
9	2		132.33	85.95
9	3		143.14	112.05
9	4		104.13	81.90
9	5		118.57	80.55
10	1	93	102.59	117.90
10	2		29.46	39.15
10	3		69.28	72.00
10	4		99.15	99.45
10	5		143.66	155.25
10	6		94.49	85.50

**Table 2: Estimated Standard Errors of Regression Coefficients  
Calculated by Alternative Procedures**

Procedure	Estimated standard Error	
	$\hat{\beta}_0$	$\hat{\beta}_1$
Ordinary least squares	10.747	0.1116
Stratified; sample weights	8.332	0.0922
Cluster; sample weights	11.121	0.0823
Stratified; unit weights	9.282	0.0963
Cluster; unit weights	13.256	0.1071

Table 3: Data for Transformed Regression Problem

Stratum Cluster	Weight	$X - \bar{X}$	$10^{-2}RWX$	RW	$\hat{v}$
1	502	-67	-195	167	6
1	502	7	25	336	6
1	502	21	68	369	-15
2	212	-48	1	1	-37
2	212	-11	4	24	-19
3	188	-30	10	-7	-20
3	188	1	5	7	-26
3	188	-17	8	-1	-35
4	190	7	4	12	5
4	190	8	4	13	-29
4	190	61	-3	38	14
5	134	-34	28	-53	34
5	134	-25	23	-51	6
5	134	0	5	-47	-3
6	189	-57	13	-20	3
6	189	6	4	11	29
6	189	25	2	20	3
7	172	45	-9	8	1
7	172	8	3	-6	-1
7	172	8	2	-6	-16
7	172	32	-5	3	-14
8	114	-6	10	-67	8
8	114	30	-22	-66	-2
8	114	-14	18	-68	28
8	114	11	-5	-67	1
8	114	42	-32	-65	5
9	193	-16	7	4	13
9	193	-6	6	9	43
9	193	21	3	22	26
9	193	-10	6	7	10
9	193	-11	6	6	35
10	114	26	-24	-90	-21
10	114	-52	63	-84	-16
10	114	-19	26	-87	-9
10	114	8	-4	-89	-6
10	114	64	65	-93	-16
10	114	-6	12	-88	3



Table 4: Efficiency of Estimated Variance under Alternative Assumptions

Procedure	Equivalent degrees of freedom	
	$V_{SRS} = V_{st}$	$V_{SRS} = 2V_{st}$
Simple random sampling	36	9
Strat. Sa., unit weights, equal var.	26.6	26.6
Strat. Sa., unequal weights, equal var.	4.8	4.8
Strat. Sa., unequal weights, $\sigma_1^2 = 0.5\sigma^2$	13.9	13.9

SELECTED BIBLIOGRAPHY OF DATA ANALYSIS FOR COMPLEX SURVEYS<sup>1</sup>

- [1] Bellhouse, D.R. (1982). Discussion provided for the Session on Data Analysis from Complex Designs. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 54-55.
- [2] Bickel, P.J. and Freedman, D.A. (1984). Asymptotic Normality and the Bootstrap in Stratified Sampling. Ann. Statist., 12, pp. 470-482.
- [3] Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimates from Complex Surveys. Intl. Statist. Review, 51, pp. 279-292.
- [4] Binder, D.A. (1982). Non-Parametric Bayesian Models for Samples from Finite Populations. J.R. Statist. Soc. B, 44, pp. 388-393.
- [5] Binder, D.A., Gratton, M., Hidiroglou, M.A., Kumar, S. and Rao, J.N.K. (1984). Analysis of Categorical Data from Surveys with Complex Designs: Some Canadian Experiences. Survey Methodology (To appear).
- [6] Brier, S.E. (1980). Analysis of Contingency Tables under Cluster Sampling. Biometrika, 67, pp. 591-596.
- [7] Cohen, J.E. (1976). The Distribution of the Chi-squared Statistics under Cluster Sampling from Contingency Tables. J. Amer. Statist. Ass., 71, pp. 665-670.
- [8] Choi, J.W. (1981). A Further Study on the Analysis of Categorical Data from Weighted Cluster Sample Survey. Proc. Amer. Statist. Ass., Section on Survey Methods Research, pp. 15-20.

---

<sup>1</sup> Prepared by the Project Team on the Analysis of Data from Complex Surveys whose members are D. Binder, M. Gratton, M. Jeays, G. Krieger, S. Kumar, D. Paton, C. Patrick and A. van Baaren.

- [9] Cowan, J. and Rinder, D.A. (1978). The Effect of a Two Stage Sample Design on Tests of Independence in a 2 by 2 Table. Survey Methodology, 4, pp. 16-28.
- [10] Ericson, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations. J. Roy. Statist. Soc. B, 31, pp. 195-223.
- [11] Fay, R.E. (1979). On Adjusting the Pearson Chi-Square Statistics for Clustered Sampling. Proc. Amer. Statist. Ass., Social Statist. Section, pp. 402-406.
- [12] Fay, R. (1981). On Jackknifing Chi-Square Test Statistics - Part I: Descriptions and Applications of the Method. Unpublished manuscript.
- [13] Fay, R. (1981). On Jackknifing Chi-Square Test Statistics - Part II: Asymptotic Theory. Unpublished manuscript.
- [14] Fay, R. (1982). Contingency Table Analysis for Complex Survey Designs: CPLX. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 44-53.
- [15] Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples. J. Amer. Statist. Ass., 75, pp. 261-268.
- [16] Fuller, W.A. (1975). Regression Analysis for Survey Data. Sankhyā C, 37, pp. 117-132.
- [17] Gross, S.J. (1980). Median Estimation in Sample Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 181-184.
- [18] Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of Categorical Data by Linear Models. Biometrics, 25, pp. 489-504.

- [19] Hidiroqlou, M.A. (1983). Approximations to the Distribution of a Sum of Weighted Chi-Square Variables. Statistics Canada, Ottawa, Ontario, Canada.
- [20] Hidiroqlou, M.A., Fuller, W.A. and Hickman, R.D. (1980). SUPER CARP. Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A.
- [21] Hidiroqlou, M.A., Fuller, W.A. and Hickman, R.D. (1980). MINI CARP. Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A.
- [22] Hidiroqlou, M.A. and Rao, J.N.K. (1983). Chi-Squared Tests for the Analysis of Three Way Contingency Tables from the Canada Health Survey. Technical report. Statistics Canada.
- [23] Holt, D. and Scott, A.J. (1981). Regression Analysis using Survey Data. The Statistician, 30, pp. 169-178.
- [24] Holt, D., Scott, A.J. and Ewings, P.O. (1980). Chi-Squared Tests with Survey Data. J.R. Statist. Soc. A, 143, pp. 302-320.
- [25] Holt, D., Smith, T.M.F. and Winter, P.D. (1980). Regression Analysis of Data from Complex Surveys. J.R. Statist. Soc. A, 143 pp. 474-487.
- [26] Imrey, P.B., Koch, G.G. and Stokes, M.E. (1981). Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part I: Historical and Methodological Review. Intl. Statist. Rev., 49, pp. 265-283 (In collaboration with J.N. Darroch, D.H. Freeman, Jr. and H.D. Tolley).
- [27] Imrey, P.B., Koch, G.G. and Stokes, M.E. (1982). Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part II: Data Analysis. Intl. Statist. Rev., 50, pp. 35-64 (In collaboration with J.N. Darroch, D.H. Freeman, Jr. and H.D. Tolley).

- [28] Imrey, P.B., Sobel, E. and Francis, M. (1980). Modeling Contingency Tables from Complex Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 212-217.
- [29] Kish, L. and Frankel, M.R. (1974). Inference from Complex Sample Surveys. J. Roy. Statist. Soc. B, 36, pp. 1-37.
- [30] Koch, G.G., Freeman, D.H., Jr. and Freeman, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. Intl. Statist. Review, 43, pp. 59-78.
- [31] Konijn, H.S. (1962). Regression Analysis in Sample Surveys. J. Amer. Statist. Ass., 57, pp. 590-606.
- [32] Landis, J.R., Lepkowski, J.M., Eklund, S.A. and Stehouwer, S.A. (1982). A Statistical Methodology for Analyzing Data from a Complex Survey. The First National Health and Nutrition Examination Survey: National Center for Health Statistics, Vital and Health Statistics, Ser. 2, No. 92. Washington, D.C.
- [33] Lepkowski, J.M. (1982). The Use of OSIRIS IV to Analyse Complex Sample Survey Data. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 38-43.
- [34] Lepkowski, J.M., Bromberg, J. and Landis, J.R. (1981). Program for the Analysis of Multivariate Categorical Data from Complex Surveys. Proc. Amer. Statist. Ass., Section on Statistical Computing, pp. 8-15.
- [35] McCarthy, P.J. (1965). Stratified Sampling and Distribution-Free Confidence Intervals for a Median. J. Amer. Statist. Ass., 60, pp. 772-783.
- [36] Nathan, G. (1969). Tests of Independence in Contingency Tables from Stratified Samples. New Developments in Survey Sampling, pp. 578-600. (N.L. Johnson, and H. Smith, eds.). Wiley: New York.

- [37] Nathan, G. (1971). A Simulation Comparison of Tests for Independence in Stratified Cluster Sampling. Bull. Int. Statist. Inst., 44(2), pp. 274-280.
- [38] Nathan, G. (1972). On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples. J. Amer. Statist. Ass., 67, pp. 917-920.
- [39] Nathan, G. (1973). Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples. National Center for Health Statistics, Vital and Health Statistics, Ser. 2, No. 53, Washington, D.C.
- [40] Nathan, G. (1975). Tables of Independence in Contingency Tables from Stratified Samples. Sankhyā C, 37, pp. 77-87.
- [41] Nathan, G. (1981). Notes on Inference Based on Data from Complex Sample Designs. Survey Methodology, 7, pp. 109-129.
- [42] Nathan, G. and Holt, D. (1980). The Effect of Survey Design on Regression Analysis. J. Roy. Statist. Soc. B, 42, pp. 377-386.
- [43] Pfefferman, D. and Nathan, G. (1977). Regression Analysis of Data from Complex Surveys. Bull. Intl. Statist. Inst., 41(3), pp. 21-42.
- [44] Rao, J.N.K. (1975). Analytic Studies of Sample Survey Data. Survey Methodology (Supplementary Issue).
- [45] Rao, J.N.K. (1983). Some Current Topics in Sample Survey Theory. Paper presented at Iowa State University Stat. Lab. 50 Anniversary Conference, June 1983.
- [46] Rao, J.N.K. and Hidiroqlou, M.A. (1981). Chi-Squared Tests for the Analysis of Categorical Data from the Canada Health Survey. Bull. Intl. Statist. Inst., 49(2), pp. 699-718.



- [47] Rao, J.N.K. and Scott, A.J. (1979). Chi-Squared Tests for Analysis of Categorical Data from Complex Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 58-66.
- [48] Rao, J.N.K. and Scott, A.J. (1981). The Analysis of Categorical Data from Complex Surveys - Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. J. Amer. Statist. Ass., 76, pp. 221-230.
- [49] Rao, J.N.K. and Scott, A.J. (1984). On Chi-Square Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. Ann. Statist., 12, pp. 46-60.
- [50] Särndal, C.E. (1982). Implications of Survey Design for Generalized Regression Estimation of Linear Functions. J. of Statist. Planning and Inference, 7, pp. 155-170.
- [51] Schuster, J.J. and Downing, D.J. (1976). Two-Way Contingency Tables for Complex Sampling Schemes. Biometrika, 63, pp. 271-276.
- [52] Scott, A.J. and Holt, D. (1982). The Effect of Two-Stage Sampling on Ordinary Least Squares Methods. J. Amer. Statist. Ass., 77, pp. 848-854.
- [53] Scott, A.J. and Rao, J.N.K. (1981). Chi-Squared Tests for Contingency Tables with Proportions Estimated from Survey Data. Current Topics in Survey Sampling. (D. Krewski, R. Platek and J.N.K. Rao, eds.)
- [54] Sedransk, J. and Meyer, J. (1978). Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling. J.R. Statist. Soc. B, 40, pp. 239-252.
- [55] Shah, B.V. (1978). SUDAAN: Survey Data Analysis Software. Proc. Amer. Statist. Ass., Section on Statistical Computing, pp. 146-151.

- [56] Shah, B.V. (1981). Development of Survey Data Analysis Software. Research Triangle Institute, Research Triangle Park, North Carolina, U.S.A.
  
- [57] Shah, B.V., Holt, M.M. and Folsom, R.E. (1977). Inference about Regression Models from Sample Survey Data. Bull. Intl. Statist. Inst., 41(3), pp. 43-57.
  
- [58] Tepping, B.J. (1968). Variance Estimation in Complex Surveys. Proc. Amer. Statist. Ass., Social Statistics Section, pp. 11-18.
  
- [59] Tomberlin, T.J. (1979). The Analysis of Contingency Tables of Data from Complex Samples. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 152-157.
  
- [60] Woodruff, R.S. (1952). Confidence Intervals for Medians and Other Position Measures. J. Amer. Statist. Ass., 47, pp. 635-646.



A Journal produced by Statistics Canada

# CONTENTS

Cost Models for Optimum Allocation in Multi-Stage Sampling WILLIAM D. KALSBECK, OPHELIA M. MENDOZA, and DAVID V. BUDESCU.....	151
Evaluation of Composite Estimation for the Canadian Labour Force Survey S. KUMAR and H. Lee.....	172
The Passenger Car Fuel Consumption Survey D. ROYCE.....	182
The Regression Estimates of Population for Sub-Provincial Areas in Canada RAVI B.P. VERMA, K.G. BASAVARAJAPPA and ROSEMARY K. BENDER.....	189
A Bibliography for Small Area Estimation.....	201

8-3200-501  
Reference No.  
Z - 079

ISSN: 0714-0045















Préparé par Statistique Canada

TABLE DES MATIÈRES

Modèles de coût pour la répartition optimale d'échantillons à plusieurs degrés	
WILLIAM D. KALSBECK, OPHELIA M. MENDOZA et DAVID V. BUDESCU.....	169
Évaluation de l'application d'estimateurs composites à l'enquête sur la population active du Canada	
S. KUMAR et H. LEE.....	196
L'enquête sur la consommation de carburant des automobiles	
D. ROYCE.....	222
Estimation par régression de la population à l'échelon infraprovincial au Canada	
RAVI B.P. VERMA, K.G. BASAVARAJAPPA et ROSEMARY K. BENDER.....	242
Une bibliographie pour l'estimation pour les petites régions.....	267





- [56] Shah, B.V. (1981). Development of Survey Data Analysis Software. Research Triangle Institute, Research Triangle Park, North Carolina, U.S.A.
- [57] Shah, B.V., Holt, M.M. et Folsom, R.C. (1977). Inference about Regression Models from Sample Survey Data. Bull. Intl. Statist. Inst., 41(3), pp. 43-57.
- [58] Tepping, B.J. (1968). Variance Estimation in Complex Surveys. Proc. Amer. Statist. Ass., Social Statistics Section, pp. 11-18.
- [59] Tomberlin, T.J. (1979). The Analysis of Contingency Tables of Data from Complex Samples. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 152-157.
- [60] Woodruff, R.S. (1952). Confidence Intervals for Medians and Other Position Measures. J. Amer. Statist. Ass., 47, pp. 635-646.

- [47] Rao, J.N.K. et Scott, A.J. (1979). Chi-Squared Tests for Analysis of Categorical Data from Complex Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 58-66.
- [48] Rao, J.N.K. et Scott, A.J. (1981). The Analysis of Categorical Data from Complex Surveys - Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. J. Amer. Statist. Ass., 76, pp. 221-230.
- [49] Rao, J.N.K. et Scott, A.J. (1984). On Chi-Square Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. Ann. Statist., 12, pp. 46-60.
- [50] Särndal, C.E. (1982). Implications of Survey Design for Generalized Regression Estimation of Linear Functions. J. of Statist. Planning and Inference, 7, pp. 155-170.
- [51] Schuster, J.J. et Downing, D.J. (1976). Two-Way Contingency Tables for Complex Sampling Schemes. Biometrika, 63, pp. 271-276.
- [52] Scott, A.J. et Holt, D. (1982). The Effect of Two-Stage Sampling on Ordinary least Squares Methods. J. Amer. Statist. Ass., 77, pp. 848-854.
- [53] Scott, A.J. et Rao, J.N.K. (1981). Chi-Squared Tests for Contingency Tables with Proportions Estimated from Survey Data. Current Topics in Survey Sampling. (édacteurs, D. Krewski, R. Platek et J.N.K. Rao.)
- [54] Sedransk, J. et Meyer, J. (1978). Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling. J.R. Statist. Soc. B, 40, pp. 239-252.
- [55] Shah, B.V. (1978). SUDAAN: Survey Data Analysis Software. Proc. Amer. Statist. Ass., Section on Statistical Computing, pp. 146-151.

- [37] Nathan, G. (1971). A Simulation Comparison of Tests for Independence in Stratified Cluster Sampling. Bull. Int. Statist. Inst., 44(2), pp. 274-280.
- [38] Nathan, G. (1972). On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples. J. Amer. Statist. Ass., 67, pp. 917-920.
- [39] Nathan, G. (1973). Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples. National Center for Health Statistics, Vital and Health Statistics, Ser. 2, No. 53, Washington, D.C.
- [40] Nathan, G. (1975). Tables of Independence in Contingency Tables from Stratified Samples. Sankhyā C, 37, pp. 77-87.
- [41] Nathan, G. (1981). L'inference statistique basée sur des plans d'échantillonnage complexes. Techniques d'enquête, 7, pp. 109-130.
- [42] Nathan, G. et Holt, D. (1980). The Effect of Survey Design on Regression Analysis. J. Roy. Statist. Soc. B, 42, pp. 377-386.
- [43] Pfefferman, D. et Nathan, G. (1977). Regression Analysis of Data from Complex Surveys. Bull. Int. Statist. Inst., 41(3), pp. 21-42.
- [44] Rao, J.N.K. (1975). Analytic Studies of Sample Survey Data. Survey Methodology (Supplementary Issue).
- [45] Rao, J.N.K. (1983). Some Current Topics in Sample Survey Theory. Exposé présenté à Iowa State University Stat. Lab. 50 Anniversary Conference, juin, 1983.
- [46] Rao, J.N.K. et Hidiroglou, M.A. (1981). Chi-Squared Tests for the Analysis of Categorical Data from the Canada Health Survey. Bull. Int. Statist. Inst., 49(2), pp. 699-718.

- [28] Jurey, P.B., Sobel, E. et Francis, M. (1980). Modeling Contingency Tables from Complex Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 212-217.
- [29] Kish, L. et Frankel, M.R. (1974). Inference from Complex Sample Surveys. J. Roy. Statist. Soc. B, 36, pp. 1-37.
- [30] Koch, G.G., Freeman, D.H., Jr. et Freeman, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. Intl. Statist. Review, 43, pp. 59-78.
- [31] Konijn, H.S. (1962). Regression Analysis in Sample Surveys. J. Amer. Statist. Ass., 57, pp. 590-606.
- [32] Landis, J.R., Lepkowski, J.M., Ekland, S.A. et Stehouwer, S.A. (1982). A Statistical Methodology for Analyzing Data from a Complex Survey. The First National Health and Nutrition Examination Survey: National Center for Health Statistics, Vital and Health Statistics, Ser. 2, No. 92. Washington, D.C.
- [33] Lepkowski, J.M. (1982). The Use of OSIRIS IV to Analyse Complex Sample Survey Data. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 38-43.
- [34] Lepkowski, J.M., Bromberg, J. et Landis, J.R. (1981). Program for the Analysis of Multivariate Categorical Data from Complex Surveys. Proc. Amer. Statist. Ass., Section on Statistical Computing, pp. 8-15.
- [35] McCarthy, P.J. (1965). Stratified Sampling and Distribution-Free Confidence Intervals for a Median. J. Amer. Statist. Ass., 60, pp. 772-783.
- [36] Nathan, G. (1969). Tests of Independence in Contingency Tables from Stratified Samples. New Developments in Survey Sampling, pp. 578-600. Redactors, N.L. Johnson, et H. Smith,). Wiley: New York.

[19] Hidiroglou, M.A. (1983). Approximations de la distribution d'une somme pondérée de variables khi-carré. Statistique Canada, Ottawa, Ontario, Canada.

[20] Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D. (1980). SUPER CARP. Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A.

[21] Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D. (1980). MINI CARP. Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A.

[22] Hidiroglou, M.A. et Rao, J.N.K. (1983). Tests khi-carré pour l'analyse d'observations classées de l'enquête santé Canada. Rapport technique. Statistique Canada.

[23] Holt, D. et Scott, A.J. (1981). Regression Analysis using Survey Data. The Statistician, 30, pp. 169-178.

[24] Holt, D., Scott, A.J. et Ewings, P.O. (1980). Chi-Squared Tests with Survey Data. J.R. Statist. Soc. A, 143, pp. 302-320.

[25] Holt, D., Smith, T.M.F. et Winter, P.D. (1980). Regression Analysis of Data from Complex Surveys. J.R. Statist. Soc. A, 143 pp. 474-487.

[26] Imrey, P.B., Koch, G.G. et Stokes, M.E. (1981). Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part I: Historical and Methodological Review. Intl. Statist. Rev., 49, pp. 265-283 (en collaboration avec J.N. Darroch, D.H. Freeman, Jr. et H.D. Tolley).

[27] Imrey, P.B., Koch, G.G. et Stokes, M.E. (1982). Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part II: Data Analysis. Intl. Statist. Rev., 50, pp. 35-64 (en collaboration avec J.N. Darroch, D.H. Freeman, Jr. et H.D. Tolley).

[9] Cowan, J. et Binder, D.A. (1978). The Effect of a Two Stage Sample Design on Tests of Independence in a 2 by 2 Table. Survey Methodology, 4, pp. 16-28.

[10] Ericson, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations. J. Roy. Statist. Soc. B, 31, pp. 195-223.

[11] Fay, R.E. (1979). On Adjusting the Pearson Chi-Square Statistics for Clustered Sampling. Proc. Amer. Statist. Ass., Social Statist. Section, pp. 402-406.

[12] Fay, R. (1981). On Jackknifing Chi-Square Test Statistics - Part I: Descriptions and Applications of the Method. Document non publié.

[13] Fay, R. (1981). On Jackknifing Chi-Square Test Statistics - Part II: Asymptotic Theory. Document non publié.

[14] Fay, R. (1982). Contingency Table Analysis for Complex Survey Designs: Appl. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 44-53.

[15] Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples. J. Amer. Statist. Ass., 75, pp. 261-268.

[16] Fuller, W.A. (1975). Regression Analysis for Survey Data. Sankhyā C, 37, pp. 117-132.

[17] Gross, S.T. (1980). Median Estimation in Sample Surveys. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 181-184.

[18] Grizzle, J.E., Starmer, C.F. et Koch, G.G. (1969). Analysis of Categorical Data by Linear Models. Biometrics, 25, pp. 489-504.



BIBLIOGRAPHIE SÉLECTIVE POUR L'ANALYSE DES DONNÉES D'ENQUÊTES COMPLEXES<sup>1</sup>

- [1] Bellhouse, D.R. (1982). Commentaires présentés à la Session on Data Analysis from Complex Designs. Proc. Amer. Statist. Ass., Section on Survey Research Methods, pp. 54-55.
- [2] Bickel, P.J. et Freedman, D.A. (1984). Asymptotic Normality and the Bootstrap in Stratified Sampling. Ann. Statist., 12, pp. 470-482.
- [3] Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimates from Complex Surveys. Intl. Statist. Review, 51, pp. 279-292.
- [4] Binder, D.A. (1982). Non-Parametric Bayesian Models for Samples from Finite Populations. J.R. Statist. Soc. B, 44, pp. 388-393.
- [5] Binder, D.A., Gratton, M., Hidiroglou, M.A., Kumar, S. et Rao, J.N.K. (1984). Analyse de données qualitatives d'enquêtes complexes: quelques expériences canadiennes. Technique d'enquête - à paraître.
- [6] Brier, S.C. (1980). Analysis of Contingency Tables under Cluster Sampling. Biometrika, 67, pp. 591-596.
- [7] Cohen, J.C. (1976). The Distribution of the Chi-squared Statistics under Cluster Sampling from Contingency Tables. J. Amer. Statist. Ass., 71, pp. 665-670.
- [8] Choi, J.W. (1981). A Further Study on the Analysis of Categorical Data from Weighted Cluster Sample Survey. Proc. Amer. Statist. Ass., Section on Survey Methods Research, pp. 15-20.

<sup>1</sup> Préparé par l'équipe de Projet de l'analyse des données d'enquêtes complexes dont les membres sont D. Binder, M. Gratton, M. Jeays, G. Kriger, S. Kumar, D. Paton, C. Patrick et A. van Baaren.

Tableau 4. Efficacité de l'estimateur de la variance dans différentes situations

Nombre équivalent de degrés de liberté		Méthode	
		$V_{EAS} = V_{st}$	$V_{EAS} = 2V_{st}$
Échantillonnage aléatoire simple (EAS)		36	9
Éch. strat., poids unitaires, var. égales		26.6	26.6
Éch. strat., poids inégaux, var. égales		4.8	4.8
Éch. strat., poids inégaux, $\sigma_1^2 = 0.5\sigma^2$		13.9	13.9

Tableau 3: Valeurs des données de régression transformées

Grappes Stratifiées	Poids	$X - \bar{X}$	$10^{-2}RWX$	RW	$\hat{v}$
1	502	-67	-195	167	6
1	502	7	25	336	6
1	502	21	68	369	-15
2	212	-48	1	1	-37
2	212	-11	4	24	-19
3	188	-30	10	-7	-20
3	188	1	5	7	-26
3	188	-17	8	-1	-35
4	190	7	4	12	3
4	190	8	4	13	-28
4	190	61	-3	38	14
5	134	-34	28	-53	34
5	134	-25	23	-51	6
5	134	0	5	-47	-3
6	189	-57	13	-20	3
6	189	6	4	11	29
6	189	25	2	20	3
7	172	45	-9	8	1
7	172	8	3	-6	-1
7	172	8	2	-6	-16
7	172	32	-5	3	-14
8	114	-6	10	-67	8
8	114	30	-22	-66	-2
8	114	-14	18	-68	28
8	114	11	-5	-67	1
8	114	42	-32	-65	5
9	193	-16	7	4	13
9	193	-6	6	9	43
9	193	21	3	22	26
9	193	-10	6	7	19
9	193	-11	6	6	35
10	114	26	-24	-90	-21
10	114	-52	63	-84	-16
10	114	-19	26	-87	-9
10	114	8	-4	-89	-6
10	114	64	65	-93	-16
10	114	-6	12	-88	3

Tableau 2. Erreurs types estimées des coefficients de régression  
calculées par différentes méthodes

Erreur type estimée	$\hat{b}_0$	$\hat{b}_1$	Méthode
0.1116	10.747		Moindres carrés ordinaires
0.0922	8.332		Echantillon stratifié; poids d'échantillonnage
0.0823	11.121		Echantillon en grappes; poids d'échantillonnage
0.0963	9.282		Echantillon stratifié; poids unitaires
0.1071	13.256		Echantillon en grappes; poids unitaires

Tableau 1. Superficie des cultures de soja  
selon deux méthodes d'observation

Hectares de soja				
Comtés	Aires	Poids	IntervIEWS (Y)	Satellite (X)
1	1	502	8.09	24.75
1	2		106.03	98.10
1	3		103.60	112.50
2	1	212	6.47	43.20
2	2		63.82	80.10
3	1	188	43.50	61.65
3	2		71.43	92.70
3	3		42.49	74.25
4	1	190	105.26	98.10
4	2		76.49	99.45
4	3		174.34	152.10
5	1	134	95.67	57.60
5	2		76.57	66.15
5	3		93.48	91.80
6	1	189	37.84	34.65
6	2		131.12	97.65
6	3		124.44	116.10
7	1	172	144.15	136.35
7	2		103.60	99.45
7	3		88.59	99.90
7	4		115.58	123.30
8	1	114	99.15	85.50
8	2		124.56	121.50
8	3		110.88	77.40
8	4		109.14	102.60
8	5		143.66	133.65
9	1	193	91.05	75.15
9	2		132.33	85.95
9	3		143.14	112.05
9	4		104.13	81.90
9	5		118.57	80.55
10	1	93	102.59	117.90
10	2		29.46	39.15
10	3		69.28	72.00
10	4		99.15	99.45
10	5		143.66	155.25
10	6		94.49	85.50

- [9] Krewski, D. et Rao, J. N. K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. Ann. Statist. 9, pp. 1010-1019.
- [10] McCarthy, P. J. (1965). Stratified sampling and distribution-free confidence intervals for the median. J. Amer. Statist. Assoc. 60, pp. 772-783.
- [11] McCarthy, P. J. (1969). Pseudo-replication: Half-samples. Rev. Int. Statist. Inst. 37, pp. 239-264.
- [12] Miller, R. G., Jr. (1974). The jackknife - a review. Biometrika 61, pp. 1-15.
- [13] Rao, J. N. K. et Wu, C. F. J. (1984). Bootstrap with stratified samples. Technical Report No. 19 of the Laboratory for Research in Statistics and Probability. Université Carleton, Ottawa, Canada.



## REMERCIEMENTS

Cette étude a été réalisée en partie grâce à une subvention accordée en vertu du Research Agreement 58-319T-1-0054X du Statistical Reporting Service du département de l'agriculture du gouvernement des États-Unis. Je remercie Nancy Hasabeinaby pour les calculs qu'elle a exécutés et Carol Francisco pour ses observations.

## BIBLIOGRAPHIE

- [1] Cochran, W. G. (1977). Sampling Techniques, 3<sup>e</sup> édition, Wiley, New York.
- [2] Efron, B. (1979). Bootstrap method: Another look at the jackknife. Ann. Statist. 7, 1-26.
- [3] Efron, B. (1981). "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika 68, pp. 589-599.

- [4] Fuller, W. A. (1975), Regression analysis for sample survey. Sankhya Series C 37, pp 117-132.

- [5] Fuller W. A. et Hidiroglou, M. A. (1978). Regression estimation after correcting for attenuation. J. Amer. Statist. Assoc. 73, pp. 99-104.

- [6] Hidiroglou, M. A., Fuller W. A. et Hickman, R. D. (1980). Super Carp, département de statistique, Iowa State University, Ames, Iowa.

- [7] Isaki, C. et Fuller, W. A. (1982). Survey design under the regression superpopulation model. J. Amer. Statist. Assoc. 77, pp. 89-96.

- [8] Kish, L. et Frankel, M. R. (1974). Inference from complex samples. J. Roy. Statist. Soc. B 36, pp. 1-22.

Liberté dans les erreurs pour le calcul de nos approximations, il est clair qu'on ne peut pas se limiter à un petit nombre de degrés de liberté, surtout si les poids sont inégaux.

Le corollaire 1 repose sur l'utilisation d'une approximation linéaire d'une fonction non linéaire des moyennes de l'échantillon pour reproduire le comportement de cette fonction non linéaire. Pour que cette approximation donne de bons résultats, la courbure de la fonction non linéaire doit être faible en comparaison de l'erreur type des moyennes de l'échantillon. Par exemple, si cette fonction est quadratique, c'est-à-dire qu'elle a la forme

$$q(\bar{y}) = \alpha_1 \bar{y} + \alpha_2 \bar{y}^2,$$

l'approximation linéaire est

$$q(\bar{y}) \doteq \alpha_1 \mu + \alpha_2 \mu^2 + (\alpha_1 + 2\alpha_2 \mu)(\bar{y} - \mu).$$

L'espérance mathématique de  $q(\bar{y})$  est

$$E\{q(\bar{y})\} = \alpha_1 \mu + \alpha_2 \mu^2 + V\{\bar{y}\}.$$

Pour que l'approximation linéaire soit bonne, il faut que la valeur de  $V\{\bar{y}\}$  (ou) celle de  $\alpha_2$  soient faibles.

En somme, on peut appliquer la théorie des grands échantillons avec assurance :

1. si on dispose d'un nombre raisonnable d'observations, c'est-à-dire un nombre suffisant pour qu'aucune observation ne soit trop éloignée des principales concentrations d'observations (c'est là une autre façon de dire que la distribution des erreurs tayloriennes est telle que leur moyenne suit approximativement une loi normale),
2. si le nombre réel de degrés de liberté dans les erreurs pour l'estimateur de la variance est raisonnable et

3. si la courbure de la fonction non linéaire des moyennes de l'échantillon est faible en comparaison de l'erreur type des moyennes de l'échantillon.

Pour établir des intervalles de confiance approximatifs, il semble raisonnable d'utiliser la loi du  $t$  de Student avec au plus  $2(n_h - 1)$  degrés de liberté. Étant donné que la variance d'une variance estimée est une fonction du quatrième moment de la population, les variances estimées sont extrêmement peu fiables. Le coefficient de variation des carrés est de  $2\frac{1}{3}$  dans le cas de la loi normale et cette valeur est beaucoup plus grande pour un grand nombre d'autres distributions souvent utilisées.

Si les variances des erreurs dans les strates sont inégales ou si des poids inégaux sont affectés aux estimations de différentes strates, la variance de l'estimateur de la variance peut être très différente de la valeur obtenue par un simple calcul du nombre de degrés de liberté des erreurs. Le tableau 4 a été construit à partir des données du tableau 1 afin d'illustrer les effets de ces facteurs sur la variance estimée. Dans la première colonne, nous supposons que la stratification n'a aucun effet; autrement dit, nous supposons que la variance dans chaque strate est égale à la variance de la population. Nous supposons que la population mère suit une loi normale, ce qui nous permet de formuler une expression explicite pour la variance de la variance. La variance de la variance estimée des erreurs dans le calcul d'une moyenne est alors proportionnelle à  $(26.6)^{-1}$  si l'échantillon est stratifié, mais elle est proportionnelle à  $36^{-1}$  si on applique les principes de l'échantillonnage aléatoire simple (EAS). Le nombre réel de degrés de liberté pour l'échantillon stratifié est d'un peu moins de vingt-sept parce que la taille des échantillons à l'intérieur des strates est inégale. Si on utilise les poids d'échantillonnage du tableau 1 et l'estimateur habituel de la variance pour échantillons stratifiés, la variance de la variance estimée est proportionnelle à  $(4.6)^{-1}$ . Ce chiffre est très bas à cause du poids élevé de la première strate. Si la variance dans la première strate est la moitié de la variance dans les autres strates, le nombre réel de degrés de liberté pour l'estimateur de la variance est de 12.4. Dans la dernière colonne du tableau 4, nous indiquons le nombre réel de degrés de liberté pour un échantillon aléatoire simple dans le cas où la variance de l'échantillon aléatoire simple est deux fois celle de l'échantillon stratifié. Ce résultat montre comment la stratification peut réduire à la fois la variance d'une moyenne estimée et la variance de la variance de cette moyenne.

Bien qu'il nous soit impossible de préciser le nombre requis de degrés de

l'observation représente moins d'un dixième du carré moyen des autres observations. En outre, l'écart quadratique pour la première observation présente un biais vers le bas parce que la méthode des moindres carrés oblique le plan d'estimation à se rapprocher d'une observation qui est séparée des autres. Par conséquent, si la variance de l'erreur sur toutes les observations est identique, la méthode Taylorienne produira une estimation de la variance du coefficient de RMX qui comporte un biais vers le bas.

Cette méthode conduit-elle à une sous-estimation de la variance pour cet échantillon? On l'ignore. Si on utilise la méthode paramétrique des moindres carrés ordinaires, on accorde beaucoup d'importance à l'observation extrême dans le calcul de l'estimation combinée de la variance de l'erreur. Il est impossible de savoir si cette méthode est exacte parce que notre estimation de la variance de l'observation isolée provient d'un estimateur à un degré de liberté. Dans ce genre de circonstances, la plupart des chercheurs spécialistes préfèrent supposer que la variance du point isolé est la même que celle des autres points au lieu d'évaluer la faible variance observée du point extrême.

Dans un modèle non paramétrique, une seule observation contient peu d'information sur la variabilité à l'intérieur de la population dans laquelle cette observation a été tirée. À toutes fins utiles, une observation qui est isolée de toutes les autres est essentiellement une seule observation. Dans la forme complète d'un modèle paramétrique, l'observation isolée est membre à part entière de l'ensemble de données parce qu'elle est censée avoir été produite par le même mécanisme générateur qui est à l'origine des autres observations. Pour des données comme celles qui figurent au tableau 3, le résultat qu'on obtient par des méthodes paramétriques est très étroitement lié à certaines hypothèses au sujet de la variance des erreurs.

Dans l'estimation de la variance, une des mesures de la grandeur numérique de l'échantillon est le nombre de degrés de liberté pour l'ensemble des données. Ainsi, par exemple, la matrice des covariances estimées pour une variable aléatoire vectorielle de dimension  $k$  est singulière, sauf si

$$\frac{1}{2} (n - 1) > k.$$

transformées contiennent les résidus obtenus dans la régression de chaque variable, sauf la première, par rapport aux éléments qui la précèdent dans le vecteur original. Les valeurs des variables de régression transformées ( $X - \bar{X}$ ,  $RW_X$ ,  $RW$ ) figurent au tableau 3. On a retenu seulement quelques chiffres pour faciliter la lecture de ce tableau.

La régression de  $Y$  par rapport à ( $1$ ,  $X - \bar{X}$ ,  $RW_X$ ,  $RW$ ) a pour résultat :

$$\hat{Y} = 95.34 + 1.085(X - \bar{X}) + 0.093 \times 10^{-2} RW_X - 0.015 RW, \\ (2.24) \quad (0.093) \quad (0.044) \quad (0.023)$$

où les erreurs types estimées ont été calculées pour un échantillon stratifié avec les poids unitaires à partir de l'expression (14). Si on utilise la méthode des moindres carrés ordinaires, on obtient

$$\hat{Y} = 95.34 + 1.085(X - \bar{X}) + 0.093 \times 10^{-2} RW_X - 0.015 RW, \\ (3.37) \quad (0.113) \quad (0.086) \quad (0.034)$$

Dans les résultats des méthodes tayloriennes, l'erreur type estimée du coefficient de  $RW_X$  est à peu près la moitié de celle qu'on obtient avec la méthode des moindres carrés ordinaires. Ce fait est attribuable à la répartition des données.

La valeur absolue de la première observation de  $RW_X$  est beaucoup plus grande que celle de toutes les autres observations. Cette seule observation représente 67% de la somme des carrés pour  $RW_X$ . Dans l'approximation taylorienne, la variance des écarts dans l'échantillon, c'est-à-dire la variance d'hi dans l'expression (14), est utilisée pour estimer la variance du paramètre d'intérêt. Les écarts de la courbe de régression,  $\hat{v}$ , figurent dans la dernière colonne du tableau 3. La valeur de  $\hat{v}$  pour la première observation est une des plus faibles dans cette colonne. Le carré moyen des résidus est de 421. Le produit  $(RW_X)(\hat{v})$  pour la première observation est égal à -113. Le produit est comparable aux résultats obtenus pour les observations 3, 33 et 36. Ainsi, la première observation est à l'origine d'environ 67% de la somme des carrés pour  $RW_X$ , mais seulement d'à peu près 15% de la somme des carrés pour  $(RW_X)(\hat{v})$ . Cette différence est due au fait que  $\hat{v}^2$  pour la première ob-



caractéristique observée est rare et se mesure avec une variable binaire (0,05), il est nécessaire de prélever un très grand échantillon (plus de 1,400 unités dans le cas d'un échantillon aléatoire simple (Cochran, 1977, p. 58)) pour une approximation basée sur la loi normale. Une variable qui suit une loi binomiale et a une faible probabilité d'être non nulle est seulement un exemple des éléments fortement asymétriques que les praticiens de l'échantillonnage rencontrent souvent dans les populations qu'ils étudient. Des grandeurs telles que le chiffre d'affaires brut d'entreprises, le nombre d'employés dans des entreprises, le nombre d'unités familiales sont des valeurs qui appartiennent à des populations asymétriques dont il faut prélever de grands échantillons avant que la distribution de la moyenne tende vers une forme gaussienne. En revanche, la distribution de la moyenne de variables telles que la taille des familles peut être approximativement normale quand la taille de l'échantillon est petite (moins de cent unités).

L'utilisation du développement de Taylor est une méthode semi-non paramétrique en ce sens que cette approximation est valable, dans les grands échantillons, si on adopte quelques hypothèses très peu restrictives au sujet de la population. Les propriétés des grands échantillons sont assurées s'il n'y a pas de points isolés dans notre espace échantillon. Cette méthode peut s'appliquer à l'échantillon, une observation ou plus est isolée du nuage de points principal. Nous examinerons maintenant le problème d'estimer la variance du vecteur de coefficients de régression utilisé pour évaluer l'effet de la pondération sur ces coefficients dans l'exemple présenté plus haut sur la culture du soja. Le vecteur original est

$$(1, X, XW, W)$$

et l'hypothèse à vérifier est que les coefficients de  $XW$  et  $W$  sont nuls. Pour illustrer les problèmes que pose l'estimation de la variance du vecteur de coefficients calculés pour l'ensemble de données sur la culture du soja, nous définissons un vecteur qui est orthogonal dans la métrique des poids unitaires. La matrice des observations des variables indépendantes



Les raisonnements présentés dans les sections précédentes sont fondés sur les propriétés des estimateurs des paramètres et de la variance dans les grands échantillons. Si on utilise la loi normale limite pour établir des intervalles de confiance, la taille d'échantillon requise pour une bonne approximation dépend de la nature de la population mère. Par exemple, si la

### 3. QU'EST-CE QU'UN GRAND ÉCHANTILLON?

complexes sur les variances estimées des coefficients de régression. Stratification et la division en grappes peuvent avoir des effets assez toujours des variances plus faibles que les poids d'échantillonnage et que la petit, mais il suffit de démontrer que les poids unitaires ne produisent pas estimateurs pondérés ont de faibles variances estimées. Cet échantillon est existe une corrélation inverse entre les poids et la variabilité observée, les moins grande que la variation à l'intérieur des grappes. Étant donné qu'il En gros, la variation de la pente d'une grappe à l'autre est relativement que l'erreur type la plus élevée correspond aux moindres carrés ordinaires. calculée pour un échantillon en grappes à partir des poids unitaires, alors fèrent. L'erreur type estimée la plus faible est associée à l'estimation Les erreurs types estimées de la pente ont toutefois un comportement différent. Les erreurs types estimées pour l'échantillon stratifié. Si on envisage l'échantillon comme un échantillon de comtés divisés en grappes, les estimations de l'erreur type de la constante sont d'environ 30% à 40% plus Si on envisage l'échantillon comme un échantillon de comtés divisés en grappes, indique un manque de convergence dans cette estimation du modèle stratifié. L'estimation de l'erreur type par la méthode des moindres carrés ordinaires proches de la droite estimée que les points correspondant aux autres strates. appartenant à la strate 1, celle qui a le poids le plus élevé, sont plus L'estimateur pondéré a la plus petite variance parce que les observations produisent un résultat convergent si on accepte le schéma de stratification. types sont des estimations. Les deux régressions pour échantillons stratifiés troisième erreurs types les moins élevées. N'oublions pas que ces erreurs unitaires et la régression des moindres carrés ordinaires ont les deuxième et dis que la régression pour échantillons stratifiés faite à partir des poids pour échantillons stratifiés produit la plus petite erreur type estimée, tan-

où les nombres entre parenthèses sont les erreurs types évaluées à partir de la matrice des covariances estimées qui est calculée à l'aide de l'équation (14). Le logiciel SUPER CARP a été utilisé pour exécuter tous les calculs nécessaires. Si on introduit les poids unitaires dans les équations (11) et (14), on obtient les résultats suivants pour l'équation de régression et les erreurs types respectivement:

$$\hat{y} = -3.927 + 1.0850x \\ (9.282) \quad (0.0963)$$

Si on calcule la variable F définie par l'équation (17), le résultat est

$$F_{23}^2 = 2.81.$$

À première vue, cette valeur est suffisamment grande pour mettre en doute l'égalité des deux coefficients. À cause de la petite taille de cet échantillon et de la structure des poids, ce test est presque l'équivalent d'une comparaison de deux droites, soit une droite pour le premier comté et une droite moyenne pour les autres comtés. Dans ce petit échantillon, les écarts de la droite du premier comté sont faibles. Par conséquent, les estimations des erreurs types des coefficients des deux variables additionnées sont peu élevées. Ce phénomène est abordé de nouveau à la section 3. Si on calcule la variable F habituellement utilisée pour la régression dans un test où on suppose que les variances des erreurs sont homogènes et qu'on fait abstraction de la stratification, on obtient

$$F_{33}^2 = 0.68.$$

Rien que cette variable ne suive pas la loi du F de Snedecor, elle nous permet de croire plus allégrement que les deux régressions pondérées faites plus haut conduisent à la même équation.

Le tableau 2 contient les erreurs types des coefficients de régression estimés par différentes méthodes. Les erreurs types estimées de la constante ont à peu près le comportement qu'on pourrait prévoir. La régression pondérée

$$\begin{pmatrix} \hat{\delta}_1' \\ \hat{\delta}_2' \end{pmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{pmatrix} \hat{Z}_{h1j} \\ \hat{Z}_{h2j} \end{pmatrix}^{-1} \begin{pmatrix} \hat{X}_{h1j} \\ \hat{X}_{h2j} \end{pmatrix},$$

et

$$\hat{V} = \begin{pmatrix} \hat{V}_{11} & \hat{V}_{12} \\ \hat{V}_{21} & \hat{V}_{22} \end{pmatrix}$$

est défini dans l'équation (14), sauf que  $\hat{Z}_{h1j}$  remplace  $\hat{X}_{h1j}$ . Tel que l'indique la notation, le critère de ce test suit approximativement la loi du  $F$  de Snedecor avec  $k$  et  $n - L - 2k$  degrés de liberté.

**Exemple 1.** Le tableau 1c contient des observations recueillies sur trente-

sept aires agricoles par le Statistical Reporting Service du département de l'Iowa en 1978. Deux séries de chiffres sur le nombre d'hectares de soja figurent dans ce tableau. La première série a été obtenue au moyen d'interviews sur place dans le cadre d'une enquête menée au mois de juin (June Enumerative Survey). La deuxième série provient d'une classification de données transmises par le satellite Landsat et analysées par un système de répartition mis au point par le Statistical Reporting Service. Ces chiffres sont tirés d'une étude qui avait pour objet la construction d'un estimateur du nombre total d'acres de soja par la méthode de régression. Nous utilisons ces données ici pour illustrer le calcul de paramètres de régression à partir de données d'enquête. L'échantillon ressemble de très près à un échantillon stratifié où les strates correspondent aux territoires numérotés dans la colonne des comtés. L'inverse de la fraction de sondage est indiqué dans la colonne des poids. L'ajustement par l'estimateur (11) d'une équation pour la régression du nombre d'hectares déclaré dans les interviews par rapport au nombre d'hectares mesuré par satellite conduit au résultat suivant:

$$\hat{y} = -11.845 + 1.1602X, \quad (8.332) \quad (0.0922)$$

issent des enquêtes est : "faut-il tenir compte des poids d'échantillonnage dans l'ajustement d'une équation de régression?" Comme pour la plupart des questions de ce genre, la réponse dépend des circonstances. Le fait que cette question soit posée indique généralement que le chercheur veut faire des inférences sur une population plus grande que la population finie qui a été échantillonnée. Cela ne signifie pas pour autant que la superpopulation d'intérêt est parfaitement définie ou définissable. Il en découle toutefois que le chercheur envisage la population finie comme étant générée par une superpopulation dans laquelle un modèle linéaire peut décrire la réalité. Une expression quantitative de l'hypothèse selon laquelle il n'est pas nécessaire de porter en compte les poids d'échantillonnage est l'hypothèse suivante concernant la superpopulation :

$$(15) \quad H_0: \theta \approx \pi = \theta(1).$$

où les  $\theta$  sont les homologues de (12) au niveau de la superpopulation et

$$\begin{aligned} \theta \approx \pi &= \left[ \begin{matrix} L \\ N_h \\ \Sigma \end{matrix} \right]_{h=1}^H \left\{ \begin{matrix} m_{hi} \\ E_{\xi} \end{matrix} \right\}_{i=1}^I \left\{ \begin{matrix} X_{hij} \pi_{hij} \\ X_{hij} \end{matrix} \right\}_{j=1}^J \left[ \begin{matrix} L \\ N_h \\ \Sigma \end{matrix} \right]_{h=1}^H \left\{ \begin{matrix} m_{hi} \\ E_{\xi} \end{matrix} \right\}_{i=1}^I \left\{ \begin{matrix} X_{hij} \pi_{hij} \\ X_{hij} \end{matrix} \right\}_{j=1}^J, \\ \theta(1) &= \left[ \begin{matrix} L \\ N_h \\ \Sigma \end{matrix} \right]_{h=1}^H \left\{ \begin{matrix} m_{hi} \\ E_{\xi} \end{matrix} \right\}_{i=1}^I \left\{ \begin{matrix} X_{hij} \tilde{\pi}_{hij} \\ X_{hij} \end{matrix} \right\}_{j=1}^J \left[ \begin{matrix} L \\ N_h \\ \Sigma \end{matrix} \right]_{h=1}^H \left\{ \begin{matrix} m_{hi} \\ E_{\xi} \end{matrix} \right\}_{i=1}^I \left\{ \begin{matrix} X_{hij} \tilde{\pi}_{hij} \\ X_{hij} \end{matrix} \right\}_{j=1}^J. \end{aligned}$$

et  $\tilde{\pi}$  est l'opérateur d'espérance mathématique par rapport à la superpopulation. Cette hypothèse est vérifiable. Il semble que le moins qu'on puisse faire, c'est d'effectuer un test de cette hypothèse si on fait une analyse non pondérée d'un échantillon tiré avec des probabilités de sélection inégales. Si l'hypothèse nulle comprend également l'hypothèse selon laquelle l'estimateur calculé en tenant compte des poids unitaires est l'estimateur à variance minimale, le test de cette hypothèse repose sur la variable

$$(17) \quad F_K^{n-L-2k} = K^{-1} \frac{\hat{\sigma}_V^2 - \hat{\sigma}_2^2}{\hat{\sigma}_2^2}.$$

quand les valeurs de  $W_{hij}$  sont proportionnelles à l'inverse des probabilités de sélection. L'erreur comprise dans  $\tilde{B}_F^W$  quand on l'utilise pour estimer  $\tilde{B}_F$  est

$$\tilde{B}_F^W - \tilde{B}_F = \left[ \sum_{h=1}^L \sum_{i=1}^n \sum_{j=1}^n W_{hij} X_{hij} X'_{hij} \right]^{-1} \sum_{h=1}^L \sum_{i=1}^n \sum_{j=1}^n W_{hij} X_{hij} V_{hij} \quad (13)$$

$$V_{hij} = Y_{hij} - X'_{hij} \tilde{B}_F.$$

En vertu du théorème 1 et du corollaire 1, un estimateur convergent de la variance de la distribution approximative de  $\tilde{B}_F^W - \tilde{B}_F$  est

$$\hat{V}\{\tilde{B}_F^W - \tilde{B}_F\} = \tilde{A}^{-1} \tilde{G} \tilde{A}^{-1}, \quad (14)$$

où

$$\begin{aligned} \tilde{A} &= \sum_{h=1}^L \sum_{i=1}^n \sum_{j=1}^n W_{hij} X_{hij} X'_{hij}, \\ \tilde{G} &= (n - 1)(n - k)^{-1} \sum_{h=1}^L \sum_{i=1}^n \sum_{j=1}^n W_{hij} V_{hij} V'_{hij}, \end{aligned}$$

$$\tilde{d}_{hij} = \sum_{j=1}^n \tilde{d}_{hij},$$

$$\tilde{d}_{hij} = W_{hij} X_{hij} V_{hij},$$

$$n = \sum_{h=1}^L \sum_{i=1}^n W_{hi},$$

$$\tilde{V}_{hij} = Y_{hij} - X'_{hij} \tilde{B}_F^W,$$

et  $\tilde{B}_F$  est l'homologue de  $\tilde{B}_F$  au niveau de la superpopulation. Cette forme particulière de l'estimateur de la variance a été proposée par Fuller (1975) et elle est utilisée par SUPER CARP.

Une des questions que les statisticiens se posent souvent quand ils ana-



une des utilisations analytiques les plus fréquentes des données d'enquête est l'ajustement d'équations de régression. On peut même exprimer la différence entre les moyennes de différents domaines sous la forme d'un coefficient de régression. Le vecteur de coefficients de régression a la forme de  $\tilde{g}(\hat{\theta})$ , variable décrite à la section précédente, mais il peut être avantageux de réécrire le vecteur  $\tilde{Y}$  défini à la section 1 en plusieurs partitions et de fournir des expressions explicites pour les coefficients de régression. L'équation de régression peut s'écrire de la manière suivante:

$$Y_{hij} = X'_{ij} \tilde{\beta} + e_{hij}, \quad (10)$$

où  $Y_{hij}$  est la variable dépendante et  $X'_{ij}$  est un vecteur de dimension  $k$  contenant les variables explicatives. L'estimateur des moindres carrés pondérés de  $\tilde{\beta}$  est

$$\tilde{\beta}_W = \left[ \sum_{h=1}^L \sum_{i=1}^I \sum_{j=1}^J n_{hi} m_{hi} X'_{ij} W_{hij} X'_{ij} \right]^{-1} \sum_{h=1}^L \sum_{i=1}^I \sum_{j=1}^J n_{hi} m_{hi} X'_{ij} W_{hij} Y_{hij}. \quad (11)$$

Les poids  $W_{hij}$  peuvent être une fonction de  $h_{ij}$ , mais nous supposons ici que les poids sont fixes en ce sens qu'ils dépendent seulement de la numérotation des unités élémentaires. On exclut ainsi (sauf à titre d'approximation) l'utilisation de poids qui varient en fonction des autres unités de l'échantillon.

Si on adopte quelques hypothèses non restrictives sur les moments de la population à laquelle la population finie appartient, on peut appliquer le théorème 1 à l'estimateur défini par la formule (11). Soit  $\pi_{hij}$  la probabilité de sélection: l'estimateur  $\tilde{\beta}_W$  est un estimateur convergent du vecteur

$$\tilde{\beta}_F = \left[ \sum_{h=1}^L \sum_{i=1}^I \sum_{j=1}^J n_{hi} m_{hi} X'_{ij} \pi_{hij} X'_{ij} \right]^{-1} \sum_{h=1}^L \sum_{i=1}^I \sum_{j=1}^J n_{hi} m_{hi} X'_{ij} \pi_{hij} Y_{hij}. \quad (12)$$

pour la population finie. Il découle de (12) que l'estimateur (11) est un estimateur convergent du coefficient de régression pour la population finie



L'estimation de la différence entre les valeurs de la moyenne par unité calculée pour les deux domaines est

$$q(\hat{\theta}) = q(\bar{y}_{...}) = \bar{y}_{-1} \bar{y}_{...3} - \bar{y}_{-1} \bar{y}_{...4} \quad (7)$$

Deux méthodes sont souvent utilisées pour calculer l'estimateur taylorien de la variance. Dans la première, l'estimateur du corollaire 1 est calculé directement à partir des matrices  $G(\hat{\theta}_I)$  et  $V\{\hat{\theta}_I - \theta_I\}$  ou  $V\{\hat{\theta}_I - \theta_{IF}\}$ . Une démarche algébriquement identique consiste à définir les observations de la manière suivante.

$$\hat{Z}(\hat{y}_{h1}, \hat{\theta}) = \hat{Z}_{h1} = G(\hat{\theta})(\hat{y}_{h1} - \bar{y}_{h..}) \quad (8)$$

et de calculer l'estimateur stratifié habituel de la variance de la moyenne par grappe pour  $\hat{Z}_{h1}$ .

$$\hat{V}\{\hat{Z}_{...}\} = \hat{V}\{G(\bar{y}_{...})\}$$

$$= \frac{1}{L} W_2^h (1 - f_h) n_h^{-1} (n_h - 1)^{-1} n_h^2 (\hat{Z}_{h1} - \hat{Z}_{h..}) (\hat{Z}_{h1} - \hat{Z}_{h..})', \quad (9)$$

où

$$\hat{Z}_{...} = \frac{1}{L} W_2^h \hat{Z}_{h..},$$

$$\hat{Z}_{h..} = n_h^{-1} \sum_{i=1}^n \hat{Z}_{hi}.$$

Par exemple, l'algorithme de calcul (9) est mis en oeuvre dans SUPER CARP. (Voir Hidiroglou et coll., 1980, p. 32.) L'analyste peut vouloir tirer des conclusions statistiques sur la population finie qui a été échantillonnée ou sur la superpopulation, quand il travaille avec des grandeurs telles que les différences entre des moyennes.

$$[\tilde{y}_{\alpha_p} - \tilde{u}_{\alpha_p}] [\tilde{y}_{\alpha_p} - \tilde{u}_{\alpha_p}]' - \partial_{\alpha_p} [\tilde{y}_{\alpha_p} - \tilde{u}_{\alpha_p}] [\tilde{y}_{\alpha_p} - \tilde{u}_{\alpha_p}]'$$

On a alors

$$[\tilde{y}_{\alpha_p} - \tilde{u}_{\alpha_p}] [\tilde{y}_{\alpha_p} - \tilde{u}_{\alpha_p}]' \xrightarrow{L} N(\tilde{0}, \tilde{I}).$$

où  $\tilde{y}_{\alpha_p} = [H(\tilde{\alpha}_p) \tilde{y}_{\alpha_p} - \tilde{\theta}_{\alpha_p} H'(\tilde{\alpha}_p)]^{-1}$ ,

et  $\tilde{u}_{\alpha_p}$  est la matrice des dérivées premières de  $\tilde{h}(\tilde{\alpha})$  par rapport à  $\tilde{\alpha}$  au point  $\tilde{\alpha}$ .

## 2. MOYENNES, QUOTIENTS ET RÉGRESSIONS

Une application élémentaire du théorème 1 est l'estimation de la moyenne par grappe et l'établissement de limites de confiance approximatives pour cette moyenne. Souvent, le paramètre clé pour l'estimateur de la moyenne est la moyenne au niveau des grappes pour la population finie, ce qui nous oblige à inclure la correction d'échantillonnage pour population finie  $(1 - f_h)$  dans l'estimateur de la variance. Une application un peu plus compliquée est l'estimation de la différence entre les valeurs de la moyenne par grappe obtenues pour deux domaines. Si nous définissons

$$\begin{aligned} y_{hij1} &= \text{observation de la caractéristique d'intérêt si l'unité hij appar-} \\ &\quad \text{tient au domaine 1} \\ &= 0 \text{ autrement,} \\ y_{hij2} &= \text{observation de la caractéristique d'intérêt si l'unité hij} \\ &\quad \text{appartient au domaine 2} \\ &= 0 \text{ autrement,} \\ y_{hij3} &= 1 \text{ si l'unité hij appartient au domaine 1} \\ &= 0 \text{ autrement et} \\ y_{hij4} &= 1 \text{ si l'unité hij appartient au domaine 2} \\ &= 0 \text{ autrement,} \end{aligned}$$

$$\frac{a\theta^T(\tilde{\theta})}{a\theta^T}$$

$q^T(\tilde{\theta})$  est le  $i$ -ième élément de  $\tilde{q}(\tilde{\theta})$  et  $\theta_j$  est le  $j$ -ième élément de  $\tilde{\theta}$ . On peut écrire

$$\begin{aligned} & [G(\tilde{\theta}_T) \tilde{Y}(\tilde{\theta}_T) - \theta_T^T G(\tilde{\theta}_T) \tilde{q}(\tilde{\theta}_T)]^{-\frac{1}{2}} [G(\tilde{\theta}_T) \tilde{Y}(\tilde{\theta}_T) - \theta_T^T G(\tilde{\theta}_T) \tilde{q}(\tilde{\theta}_T)]^{\frac{1}{2}} \xrightarrow{L} N(0, I). \\ & [G(\tilde{\theta}_T) \tilde{Y}(\tilde{\theta}_T) - \theta_T^T G(\tilde{\theta}_T) \tilde{q}(\tilde{\theta}_T)]^{-\frac{1}{2}} [G(\tilde{\theta}_T) \tilde{Y}(\tilde{\theta}_T) - \theta_T^T G(\tilde{\theta}_T) \tilde{q}(\tilde{\theta}_T)]^{\frac{1}{2}} \xrightarrow{L} N(0, I). \end{aligned}$$

Le corollaire 1 s'applique à l'estimateur taylorien de la variance de la distribution approximative de  $\tilde{q}(\tilde{\theta}_T) - q(\tilde{\theta}_T)$ . Il est également possible d'évaluer la variance à l'aide d'estimateurs bien définis pour des échantillons répétés. Quelques-unes des techniques utilisées sont la méthode des duplications successives par blocs (balanced repeated replication: voir McCarthy (1969)), les méthodes jackknife (voir Miller (1974)) et les méthodes bootstrap (voir Efron (1979, 1981)). Rien que ces méthodes puissent être adaptées à la structure de l'échantillon, cette adaptation n'est pas toujours directe (voir Rao et Wu (1983)).

Une classe de fonctions continues de  $\tilde{\theta}$  qui mérite une attention spéciale est celle qu'on obtient quand  $\tilde{\theta}$  est la variable dépendante dans un ajustement par la méthode des moindres carrés généralisés.

**Corollaire 2.** Supposons que les hypothèses du théorème 1 sont justes. Définissons  $\theta$  de manière à satisfaire la condition

$$\tilde{\theta} = h(\tilde{\alpha}),$$

où  $\alpha$  est un vecteur de dimension  $\tilde{k}$  ( $k \leq p$ ),  $h(\alpha)$  est une fonction continue de  $\alpha$  et il existe des dérivées premières et seconde pour tout  $\alpha$  dans une sphère ouverte contenant la vraie valeur  $\alpha^*$  pour tout  $r$ . Définissons l'espace des paramètres de  $\alpha$  comme un sous-ensemble borné ouvert de l'espace euclidien à  $k$  dimensions. Soit  $\tilde{\alpha}^*$  le vecteur qui minimise

$$\sup_L \left[ \sum_{t=1}^h W_2^t u_{-1}^t - W_2^h u_{-2}^h \right] \longrightarrow 0.$$

à mesure que  $\tau \longrightarrow \infty$ , ou  $W^{\tau h}$  est une matrice triangulaire de poids. On peut

écrite

$$\begin{aligned} [ \tilde{V}_{\tilde{\theta}}^{\tau} - \tilde{\theta}_{\tau}^{\tau} ]^{-\frac{1}{2}} &\xrightarrow{L} N(\tilde{0}, \tilde{I}), \\ [ \tilde{V}_{\tilde{\theta}}^{\tau} - \tilde{\theta}_{\tau}^{\tau} ]^{-\frac{1}{2}} &\xrightarrow{L} N(\tilde{0}, \tilde{I}), \end{aligned}$$

où

$$\begin{aligned} [ \tilde{V}_{\tilde{\theta}}^{\tau} - \tilde{\theta}_{\tau}^{\tau} ]^{-\frac{1}{2}} &= \frac{1}{L} W_2^{\tau} (1 - F^{\tau h})^{-\frac{1}{2}} \tilde{V}_{\tilde{\theta}}^{\tau h}, \\ [ \tilde{V}_{\tilde{\theta}}^{\tau} - \tilde{\theta}_{\tau}^{\tau} ]^{-\frac{1}{2}} &= \frac{1}{L} W_2^{\tau} W_{-1}^{\tau h} \tilde{V}_{\tilde{\theta}}^{\tau h}. \end{aligned}$$

$$\begin{aligned} \tilde{V}_{\tilde{\theta}}^{\tau h} &= (n^{\tau h} - 1)^{-1} \sum_{i=1}^{\tau h} (\tilde{V}_{\tilde{\theta}}^{\tau h}) (\tilde{V}_{\tilde{\theta}}^{\tau h})' - \tilde{V}_{\tilde{\theta}}^{\tau h}, \\ \tilde{V}_{\tilde{\theta}}^{\tau h} &= n^{\tau h} \sum_{i=1}^{\tau h} \tilde{V}_{\tilde{\theta}}^{\tau h}. \end{aligned}$$

La démonstration de ce théorème découle des théorèmes 1 et 2 de Fuller (1975) et elle est applicable aux échantillons à plusieurs degrés. (Voir aussi Krewski et Rao (1981) et Isaki et Fuller (1982)).

La plupart des cas que nous examinerons portent sur des fonctions continues de  $\tilde{\theta}$ .

**Corollaire 1.** Supposons que les hypothèses du théorème 1 sont justes. Soit  $\tilde{q}(\tilde{\theta})$  une fonction à valeurs vectorielles qui dépend de  $\tilde{\theta}$ , où  $\tilde{q}(\tilde{\theta})$  est continue et admet des dérivées premières continues pour  $\tilde{\theta}$  dans la sphère  $|\tilde{\theta} - \tilde{\theta}^{\tau}| \leq \delta$  pour tout  $\tau$ , où  $\delta > 0$  est une valeur fixe. Soit  $\tilde{G}(\tilde{\theta})$  une matrice non singulière contenant les dérivées premières de  $\tilde{q}(\tilde{\theta})$ , où le  $ij$ -ième élément de  $\tilde{G}(\tilde{\theta})$  est

comme un échantillon aléatoire de taille  $N^{rh} \geq N^{r-1,h}$  choisis à partir d'une population infinie à  $p$  dimensions admettant des moments absolus d'ordre  $2 + \delta$ , où  $\delta > 0$ , qui sont bornés en vertu de l'inégalité  $M_\delta < \infty$ . Soit  $\Sigma^{rh}$  la matrice des covariances de la  $rh$ ème population infinie. Soit  $L^r \geq L^{r-1}$  le nombre de strates dans chaque population finie et supposons qu'un échantillon aléatoire simple de  $n^{rh}$  unités ( $n^{rh} \geq 2$  et  $n^{rh} \geq n^{r-1,h}$ ) est tiré dans la  $h$ ème strate, Soit  $f^{rh} = N^{rh} n^{rh}$  une matrice triangulaire dans lequel

$$0 \leq f^{rh} < M^{fu} < 1,$$

où  $M^{fu}$  est un nombre fixe. Soit  $Y^{rh}$ . le total correspondant à la  $i$ ème grappe tirée dans la  $h$ ème strate de la  $r$ ème population et définissons

$$\tilde{x}^{rh} = \frac{1}{L^{rh}} \sum_{i=1}^{n^{rh}} w^{rh} x^{rh}_i, \quad y^{rh} = \frac{1}{N^{rh}} \sum_{i=1}^{n^{rh}} w^{rh} y^{rh}_i,$$

$$\tilde{x}^{rh} = \frac{1}{L^{rh}} \sum_{i=1}^{n^{rh}} w^{rh} x^{rh}_i, \quad y^{rh} = \frac{1}{N^{rh}} \sum_{i=1}^{n^{rh}} w^{rh} y^{rh}_i,$$

$$\tilde{x}^{rh} = \frac{1}{L^{rh}} \sum_{i=1}^{n^{rh}} w^{rh} x^{rh}_i, \quad y^{rh} = \frac{1}{N^{rh}} \sum_{i=1}^{n^{rh}} w^{rh} y^{rh}_i,$$

où  $\tilde{x}^{rh}$  est un paramètre de la population finie et  $y^{rh}$  est la moyenne de la population infinie utilisée pour produire la  $h$ ème strate de la population finie. Nous supposons également que

$$0 < M^{SL} < \left| \frac{1}{L^{rh}} \sum_{i=1}^{n^{rh}} w^{rh} x^{rh}_i \right| < M^{SU} < \infty,$$

où les bornes inférieure et supérieure  $M$  sont des nombres fixes: on suppose que

$$x^{rh} = \frac{1}{L^{rh}} \sum_{i=1}^{n^{rh}} w^{rh} x^{rh}_i \rightarrow \infty.$$

locallement d'une fonction linéaire du vecteur des moyennes de chaque grappe

$$\hat{\theta} = \sum_{h=1}^L w_h^{-1} \sum_{i=1}^n y_i \sim y_h. \quad (4)$$

où les  $w_h$  sont des poids fixes. Souvent, ces poids valent

$$w_h = N_h^{-1} \quad (5)$$

où  $N_h$  est le nombre de grappes dans la  $h^{\text{ème}}$  strate et  $N$  est le nombre total de grappes au niveau de la population. Pour les poids définis dans la formule (5), la fonction linéaire en (4) est l'estimateur centré habituellement utilisé pour calculer la moyenne par grappe pour une population finie. Un autre groupe de poids qui présente beaucoup d'intérêt est l'ensemble de poids unitaires

$$w_h = n^{-1} n_h. \quad (6)$$

Notre modèle nous permet d'examiner des fonctions de la moyenne par unité élémentaire. L'estimateur habituel de la moyenne d'une des variables  $Y$  au niveau des unités élémentaires est le rapport entre la moyenne de la variable  $Y$  au niveau des grappes et le nombre moyen d'unités par grappe. Le nombre moyen d'unités par grappe est la moyenne au niveau des grappes d'une variable  $Y$  qui est identiquement égale à l'unité.

On peut facilement étendre notre modèle pour inclure diverses formes de sous-échantillonnage à l'intérieur des grappes. Étant donné que ce genre d'extension n'augmente guère la généralité du développement et qu'elle compliquerait beaucoup les notations, nous limiterons notre description à l'échantillonnage à un degré à l'intérieur des strates.

Notre analyse repose fondamentalement sur le théorème central limite suivant pour les échantillons tirés dans une population finie.

**Théorème 1.** Soit  $\{\xi_r: r = 1, 2, \dots\}$  une série de populations finies stratifiées. Définissons la population de la  $h^{\text{ème}}$  strate de la  $r^{\text{ème}}$  population



# APPLICATION DE LA MÉTHODE DES MOINDRES CARRÉS ET DE TECHNIQUES CONNEXES AUX PLANS DE SONDAGE COMPLEXES

Wayne A. Fuller<sup>1</sup>

## 1. INTRODUCTION ET DESCRIPTION DU MODÈLE

Supposons qu'un échantillon de grappes d'unités élémentaires est prélevé dans une population finie divisée en  $L$  strates. Le nombre total de grappes (unités primaires d'échantillonnage) dans l'échantillon,  $n$ , est défini par l'expression

$$n = \sum_{h=1}^L n_h, \quad (1)$$

où  $n_h \geq 2$  est le nombre de grappes tirées de la  $h^{\text{ième}}$  strate. Un vecteur colonne de caractéristiques

$$Y_{hij} = (Y_{hij1}, Y_{hij2}, \dots, Y_{hijp})' \quad (2)$$

est observé pour la  $j^{\text{ième}}$  unité élémentaire appartenant à la  $i^{\text{ième}}$  grappe de la  $h^{\text{ième}}$  strate. La forme du vecteur  $Y_{hij}$  est assez générale. Ainsi, certains éléments de ce vecteur peuvent être des puissances de produits d'autres composantes du même vecteur. Un élément peut aussi, et ce sera souvent le cas, être identiquement égal à l'unité. Les totaux pour chaque grappe sont calculés à l'aide de la formule

$$Y_{hi} = \sum_{j=1}^{m_{hi}} Y_{hij}, \quad (3)$$

où  $m_{hi}$  est le nombre d'unités élémentaires dans la  $hi^{\text{ième}}$  grappe. Nous voulons nous pencher sur le comportement de fonctions continues

- [21] Rao, J.N.K. et Scott, A.J. (1984). On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. Annals of Statistics 12, pp. 46-60.
- [22] Rosenbaum, P.R.R. et Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies. Biometrika 70, pp. 41-55.
- [23] Rubin, D.B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. Annals of Statistics 6, pp. 34-58.
- [24] Rubin, D.B. (1983). Comment: Probabilities of Selection and Their Role for Bayesian Modeling in Sample Surveys. Journal of the American Statistical Association 78, pp. 803-805.
- [25] Särndal, C.-E. (1978). Design-Based and Model-Based Inference in Survey Sampling. Scandinavian Journal of Statistics 5, pp. 27-52.
- [26] U.S. Bureau of the Census (1982). Preliminary Evaluation Results. Memorandum No. 31: Evaluating the Public Information Campaign for the 1980 Census - Results of the 1980 KAP Survey. Document rédigé par Jeffrey C. Moore, Washington, D.C.

- [11] Fay, R.E. (1984). A Jackknifed Chi-Square Test for Complex Samples. A paraitre dans la revue Journal of the American Statistical Association.
- [12] Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples. Journal of the American Statistical Association 75, pp. 261-268.
- [13] Fuller, W.A. (1975). Regression Analysis for Sample Survey. Sankhya C 37, pp. 117-132.
- [14] Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). Sample Survey Methods and Theory, Volumes I and II. New York: John Wiley.
- [15] Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. Journal of the American Statistical Association 78, pp. 776-793.
- [16] Hidiroglou, M.A., Fuller W.A. et Hickman, R.D. (1978). Super Camp (3e édition). Ames, IO: Statistical Laboratory, Iowa State University.
- [17] Kish, L. (1965). Survey Sampling. New York: John Wiley.
- [18] Kish, L. et Frankel, M.R. (1974). Inference from Complex Samples. Journal of the Royal Statistical Society, Ser. B 36, pp. 1-37.
- [19] Koch, G.G., Freeman, D.H. et Freeman, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Samples. International Statistical Review 43, pp. 59-78.
- [20] Rao, J.N.K. et Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. Journal of the American Statistical Association 76, pp. 221-230.

- [2] Bishop, Y.M.M., Fienberg, S.E. et Holland, P.W. (1975). Discrete Multivariate Analysis. Cambridge, MA: MIT Press.
- [3] Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: John Wiley.
- [4] Cochran, W.G. (1977). Sampling Techniques (3e édition). New York: John Wiley.
- [5] Dippo, C.S., Fay, R.E. et Morgantstein, D.H. (1984). Computing Variances from Complex Samples with Replicate Weights. Communication présentée à la réunion annuelle de l'American Statistical Association, section des techniques d'enquête.
- [6] DuMouchel, W.H. et Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association 78, pp. 535-543.
- [7] Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- [8] Fay, R.F. (1982). Contingency Tables for Complex Designs: CPLX. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 44-53.
- [9] Fay, R.E. (1983). CPLX - Contingency Tables Analysis for Complex Sample Designs, Program Documentation. Document non publié, Washington, D.C.: U.S. Bureau of the Census.
- [10] Fay, R.E. (1984). Some Properties of Estimates of Variance Based on Replication Methods. Communication présentée à la réunion annuelle de l'American Statistical Association, section sur les techniques d'enquête.

directement aux estimations pondérées d'une enquête. Cette méthode ressemble beaucoup aux procédés élaborés par Rao et Scott. Aucune comparaison exhaustive des avantages relatifs du test par la méthode du jackknife et des tests proposés par Rao et Scott n'a été faite jusqu'à présent, mais il semble à première vue qu'ils soient tous bons et que ni l'un, ni l'autre ne soit supérieur. (D'autres observations à ce sujet sont présentées dans [11].)

Les tests par la méthode du jackknife semblent toutefois un peu plus faciles à appliquer, surtout quand un tableau contient un grand nombre de cases. Un programme écrit en FORTRAN et intitulé CPLX ([8]) le décrit et [9] en explique le mode d'utilisation) comprend des tests par le jackknife pour l'ajustement de modèles factoriels log-linéaires aux classements recoupés et est maintenant à la disposition du public. Ce programme calcule également des erreurs types pour les paramètres de modèles log-linéaires par une méthode de duplication, ce qui permet de résoudre le premier des trois problèmes d'inférence examinés plus haut. CPLX est très pratique quand les autres variances des paramètres d'une enquête sont aussi estimées à l'aide de techniques de duplication, comme celles fondées sur les poids d'échantillonnage artificiels décrits à la section précédente, mais ce n'est pas là une condition obligatoire pour l'utilisation de ce programme et beaucoup de spécialistes à l'intérieur et à l'extérieur du Census Bureau l'ont appliqué à divers types d'analyses.

L'auteur espère pouvoir éventuellement intégrer la méthodologie de Rao et Scott dans un programme comme CPLX afin que les utilisateurs de données puissent aussi disposer de cette technique. Pour le moment, toutefois, la version actuelle de CPLX devrait aider les chercheurs qui veulent appliquer des procédés d'inférence basés sur le plan de sondage aux données d'enquête.

## BIBLIOGRAPHIE

- [1] Binder, D.A. (1983). On the variances of asymptotically normal estimators from Complex Samples. International Statistical Review 51, pp. 279-292.

recoupés complets de données qualitatives, comprennent un grand nombre de paramètres. Les trois problèmes d'inférence les plus fréquents sont :

1. Le calcul de l'erreur type et des intervalles de confiance pour chaque paramètre estimé.
2. La vérification de l'importance de l'effet d'ensembles précis de paramètres sur la qualité de l'ajustement d'un modèle.
3. La vérification de la validité globale de l'ajustement du modèle.

Dans un échantillon aléatoire simple, les principes classiques de la méthode de du maximum de vraisemblance offrent une solution à ces problèmes, quoique le test du khi-carré de Pearson soit, à juste titre, plus répandu que le test du khi-carré lié au rapport de vraisemblance pour résoudre le troisième problème.

Koch, Freeman et Freeman [9] ont appliqué la méthode des moindres carrés pondérés (MCP) aux échantillons complexes pour obtenir une solution à chacun des trois principaux problèmes d'inférence décrits plus haut. Cette méthode s'est avérée très utile en général, mais, dans certains cas, elle est limitée par la nécessité de produire des estimations très précises corrigées en fonction du plan de sondage pour la covariance des estimations calculées à partir de l'échantillon avant que ses propriétés asymptotiques puissent reproduire les résultats des MCP. (Pour d'autres observations sur les limites des MCP, voir [8] et [11].)

Felleci [12] a conçu un des premiers tests pour évaluer les résultats de la méthode des MCP dans des cas particuliers. Plus récemment, Rao et Scott [20], [21] ont formulé et étendu un ensemble de techniques connexes pour des tests d'une classe générale de modèles comprenant les modèles log-linéaires. Statistique Canada a participé activement à l'élaboration de ces méthodes.

Un test moins bien connu unit la technique du jackknife et la variable khi-carré [11] pour résoudre le problème général des tests d'hypothèses modifiées en fonction du plan de sondage. Ce test repose sur la duplication de l'échantillon, et l'équation (4.4) et une expression semblable pour l'approximation du biais d'ordre premier (comme d'habitude dans le jackknife) sont utilisées pour faire des inférences approximatives concernant la distribution postulée par l'hypothèse nulle dans les tests habituels du khi-carré appliqués



La modularité de ces trois phases est un des grands avantages de cette technique; des programmes généraux peuvent servir à l'exécution des phases 1 et 2 ou des programmes spéciaux peuvent être mis au point si des problèmes particuliers se posent. Pour une même enquête, il est nécessaire d'exécuter les phases 1 et 2 seulement une fois. Les programmes conçus pour la phase 3 ne renferment pas de corrections en fonction du plan de sondage ou de l'estimateur et peuvent donc être utilisés au besoin par tout utilisateur qui dispose des poids artificiels  $W_{IT}$  produits à la deuxième phase.

La plupart des applications de cette méthode au Censur Bureau ont porté sur l'estimation de la variance de caractéristiques fondamentales telles que des moyennes, des totaux ou des proportions dans des enquêtes, mais l'équation (4.4) est également utile pour des travaux analytiques. Cette méthode permet de représenter tous les effets des plans de sondage et des estimateurs complexes, mais, en pratique, l'application des techniques de linéarisation se limite souvent à des cas simples et assez fréquents. Par ailleurs, bien qu'on puisse élaborer des logiciels pour appliquer la linéarisation aux méthodes analytiques répandues, telles que la régression linéaire, les modèles log-linéaires, les modèles linéaires généralisés, etc., la formule (4.4) permet de calculer des variances dans des modèles analytiques plus spécialisés pour lesquels il n'existe pas de programmes de linéarisation, car (4.4) nécessite seulement l'exécution d'algorithmes complets en fonction des nouvelles estimations produites à partir des poids artificiels.

## 5. MÉTHODES D'INFÉRENCE BASÉES SUR LE PLAN DE SONDAGE POUR LES MODÈLES LOG-LINÉAIRES

Les modèles log-linéaires, qui expriment le logarithme des fréquences théoriques des valeurs d'une variable qualitative sous la forme d'une fonction linéaire de paramètres inconnus, comprennent les modèles factoriels pour les classements recoups de données qualitatives et les modèles logistiques, qui définissent la relation entre une variable qualitative dépendante ou plus et une combinaison quelconque de variables explicatives qualitatives et continues. Bishop, Fienberg et Holland [2] ont publié un des premiers ouvrages sur ce domaine qui se développe très rapidement.

Beaucoup de modèles log-linéaires, surtout ceux conçus pour des classements

$$\text{Var}(\hat{X}_0) = \sum_{r=1}^R d_r (\hat{X}_r - \hat{X}_0)^2 \quad (4.3)$$

pour une valeur prédéterminée de  $d_r$  qui est indépendante du choix de la variance  $X$ . Par exemple, dans un échantillon séparé en deux parties égales équilibrées, une manière simplifiée d'estimer la variance, abstraction faite des effets des estimations complexes qui peuvent influencer sur les poids  $W_{i0}$ , consisterait à fixer les poids  $W_{ir}$  égaux soit à  $2W_{i0}$ , soit à  $\eta$ , selon que l'unité  $i$  fait partie ou non du demi-échantillon, et de fixer  $d_r = 1/R$  pour chaque  $r$ . Plus généralement, pour une fonction lisse  $S$  des estimations pondérées de  $X$  à l'échelle de la population  $X_0(1), \dots, X_0(k)$ , qui ont toutes la forme (4.1),

$$\text{Var}(\hat{X}_0) = \sum_{r=1}^R d_r \{S(\hat{X}_r(1), \dots, \hat{X}_r(k)) - S(\hat{X}_0(1), \dots, \hat{X}_0(k))\}^2 \quad (4.4)$$

Dans l'équation (4.4),  $S$  peut être un des estimateurs extrêmement complexes qui sont souvent utilisés dans les enquêtes et peuvent inclure des corrections pour tenir compte des cas de non-interview et des estimations directes ou indirectes par le quotient. En outre, si ces formes d'estimation complexe sont intégrées dans les poids  $W_{ir}$ , elles peuvent également être incluses dans le calcul des valeurs de  $W_{ir}$ . Bref, le calcul de la variance dans cette méthode comprend trois étapes ou phases distinctes :

1. Détermination des poids d'échantillonnage artificiels de base  $W_{ir}^*$  pour le schéma de pondération simple sans biais des données (schéma de Horwitz-Thompson) qui correspond aux poids de base  $W_{i0}^*$ .
2. Calcul des poids d'échantillonnage artificiels finals,  $W_{ir}$ , par l'application des mêmes rectifications pour les cas de non-interview et les estimations par le quotient,  $W_{ir}^*$ , que dans les premières méthodes d'estimations utilisées pour obtenir les  $W_{i0}^*$  à partir des  $W_{i0}^*$ .
3. Estimation de la variance de paramètres simples ou complexes à l'aide de l'équation (4.4).

teur (qui peut être complexe), l'algorithme de cette méthode représente la variance sous la forme de données qui sont associées au fichier des données d'enquête plutôt que comme un ensemble de formules de variance (qui peuvent être complexes) pour lesquelles il faut composer un programme informatique. Les méthodes de duplication bien connues, comme la séparation de l'échantillon en deux parties égales et équilibrées et le jackknife, peuvent être formulées en fonction de poids d'échantillonnage artificiels, mais l'algorithme mentionné plus haut permet également de reproduire un gamme très étendue de plans de ré-échantillonnage (voir [71]). Le document [10] montre qu'il existe un plan de ré-échantillonnage (plus exactement, un nombre infini de plans de ré-échantillonnage) qui correspond essentiellement à n'importe lequel des estimateurs de la variance bien connus pour les estimations de totaux dans une population. De cette manière, on peut, par exemple, reproduire les expressions de la variance pour les plans de sondage à plusieurs degrés, les estimateurs de Yates-Grundy et ainsi de suite. La représentation des relations de variance complexes sous la forme de données rend le calcul des variances accessible à un plus grand nombre d'utilisateurs de données.

Dans un grand nombre d'enquêtes, des poids  $W_{i0}$  sont affectés à chaque unité  $i$ , de sorte que, pour toute variable  $X_i$ , l'estimation d'un total correspond à la somme pondérée du produit de la variable et des poids des unités d'échantillonnage:

$$\hat{X}_0 = \sum_{i=1}^I W_{i0} X_{i0} \quad (4.1)$$

Le résultat de la méthode des poids d'échantillonnage artificiels est un nouvel ensemble de poids  $W_{ir}$ ,  $r = 1, \dots, R$ , pour chaque unité d'échantillonnage  $i$ , à partir desquels on peut calculer une autre estimation d'un total:

$$\hat{X}_r = \sum_{i=1}^I W_{ir} X_{ir} \quad (4.2)$$

L'estimateur de la variance est

de linéarisation. L'avantage de ce type de logiciel est qu'il permettrait d'explorer les méthodes de duplication élaborées pour quelques-unes de nos enquêtes et de mieux tenir compte des effets des estimateurs complexes que les programmes fondés sur les procédés de linéarisation.

#### 4. CALCUL DES VARIANCES CORRIGÉES POUR TENIR COMPTE DU PLAN DE SONDAGE À L'AIDE DE POIDS D'ÉCHANTILLONNAGE ARTIFICIELS

Les méthodes de duplication telles que le jackknife, la séparation en deux parties égales et le bootstrap représentent les principales solutions de recours aux techniques de linéarisation pour calculer des estimations de variance. Les paramètres non linéaires en fonction du plan de sondage. Kish et Taniguchi [18] ont publié un premier compte rendu des applications des méthodes de duplication à ce problème, et beaucoup de travaux ont été réalisés par la suite.

L'utilisation des méthodes de duplication pour l'estimation de la variance a eu des hauts et des bas. La linéarisation est une technique puissante, bien entendu, et les propriétés mises en évidence par Binder [1] favorisent son application à un grand éventail de modèles analytiques. Toutefois, les enquêtes du Census Bureau reposent en général sur des estimateurs assez complexes, et l'évaluation de l'effet total de la variance d'échantillonnage de ces estimateurs s'est souvent avérée une tâche qui nécessite beaucoup de temps de la part de spécialistes en statistique et, surtout, en programmation informatique. Récemment, les calculs de la variance dans diverses enquêtes ont pu être exécutés par des méthodes de duplication fondées sur des "poids d'échantillonnage artificiels" (replicate weights). Cette méthode est essentielle pour un procédé général pour calculer la variance d'un grand nombre de statistiques dans les enquêtes et simplifier l'estimation de la variance des paramètres analytiques complexes.

L'utilisation de poids d'échantillonnage artificiels n'est pas une découverte récente. Quelques-unes des premières applications de cette technique sont résumées en [5], qui décrit également l'expérience acquise dans ce domaine par le U.S. Bureau of Labor Statistics, le Bureau of the Census et Westat, Inc. Pour un plan de sondage (qui peut être complexe) et un estima-

au Censu Bureau semble donc concorder avec le critère établi par DuMouchel et Duncan: il faut choisir un estimateur efficace (et simple) si le modèle hypothétique est vraisemblable ou un estimateur convergent si le modèle est inacceptable. On utilise le plus souvent des méthodes basées sur le plan de sondage pour faire des inférences à l'échelle nationale, tandis que les procédés basés sur un modèle servent d'outils pour des analyses moins strictes ou des analyses dans lesquelles on espère qu'un modèle théorique est exact (quand des données sont manquantes).

### 3. MÉTHODE UTILISÉE AU CENSUS RURAL POUR L'INFÉRENCE EN FONCTION DU PLAN DE SONDAGE DANS LES RÉGRESSIONS LINÉAIRES

En général, dans les travaux statistiques, la régression linéaire est sans doute la technique analytique la plus répandue. La plupart des données recueillies par le Censu Bureau, en particulier celles provenant d'études démographiques qui révèlent des caractéristiques des personnes ou des logements, sont qualitatives. Le Censu Bureau utilise donc la régression linéaire, sous quelque forme que ce soit, relativement moins qu'ailleurs.

Fuller [13] a élaboré les notions de base pour l'application des techniques d'inférence à la régression linéaire à partir de méthodes (de linéarisation) fondées sur le développement de Taylor. Ses résultats sont intégrés au programme informatique SUPER CARP [16], dont la mise au point a été en partie financée par le U.S. Bureau of the Census. Nous pouvons affirmer que nous avons utilisé ce logiciel avec succès, mais nous l'avons appliqué à seulement un petit nombre de problèmes jusqu'à présent. Le document de travail de Moore [26] offre probablement l'illustration la plus accessible de l'utilisation de SUPER CARP au Censu Bureau.

Dans la section suivante, on explique la manière de définir les méthodes de duplication en fonction de poids d'échantillonnage artificiels (replicates weights) et on présente aussi des réflexions préliminaires sur la conception de logiciels pour exécuter les calculs nécessaires, quoiqu'on n'ait pas encore essayé de les mettre en oeuvre. Cette méthode n'est pas très différente de celle utilisée dans SUPER CARP, mais les méthodes de duplication aboutissent souvent à des valeurs un peu plus élevées des erreurs types que les techniques



Le comportement prévu dans le modèle théorique, toute dépendance entre les observations causée par l'effet de grappe (quand elle est assez prononcée) contredit automatiquement toute hypothèse d'indépendance des erreurs qui peut faire partie d'un modèle trop simplifié. Ainsi, les modèles qui font abstraction des effets de grappe quand on sait que ceux-ci existent sont forcément des représentations inexactes des données en question.

Les techniques d'inférence basées sur le plan de sondage constituent la norme au U.S. Bureau of the Census; toutefois, en pratique, les deux formes d'inférence sont unies dans l'ajustement de modèles. La plupart des chercheurs ont tendance à appliquer des techniques basées sur le plan de sondage pour faire des inférences concernant des relations à l'échelle nationale à partir d'échantillons complexes. Quand la variation des poids d'échantillon-nage est modérée ou nulle et qu'on peut supposer que les effets de grappe sont faibles, les méthodes d'inférence basées sur un modèle semblent être bien acceptées. L'intérêt manifesté à l'égard des procédés basés sur un modèle dans ces cas découle sans doute plus d'un choix pratique que d'un point de vue philosophique: les méthodes basées sur un modèle sont plus accessibles et mieux connues que les techniques correspondantes basées sur le plan de sondage. (L'auteur a observé des cas où la variation des poids et les effets de grappe étaient tels que les méthodes basées sur le plan de sondage reproduisaient exactement les mêmes conclusions que celles basées sur un modèle, ce qui justifie l'utilisation de techniques basées sur un modèle dans ce genre de conditions. Toutefois, quand les poids d'échantillonnage varient beaucoup ou qu'il existe des effets de grappes dans certaines caractéristiques, il est facile de trouver des exemples où ces deux formes d'inférence entraînent des résultats très différents et où les méthodes d'inférence basées sur un modèle sont fort douteuses.)

Dans certains domaines d'activité précis du Census Bureau, il semble que ce soit exclusivement les méthodes basées sur un modèle qui sont utilisées. En particulier, les méthodes d'imputation de données manquantes, dont certaines reposent sur des modèles paramétriques explicites, sont caractérisées par l'absence d'un rôle pour les poids renfermés dans le plan de sondage. Un autre domaine d'étude, celui des estimations relatives aux petites régions ou aux petits secteurs, comporte souvent un mélange de procédés d'inférence basés sur le plan de sondage et un modèle. L'application des techniques d'inférence



souligné le fait qu'une interprétation bayésienne complète des données observées repose non seulement sur les relations fonctionnelles et distribuées telles que (2.1) à l'échelle de la population, mais aussi sur le processus par lequel les observations de l'échantillon sont recueillies. (Dans un plan de sondage randomisé, la "propension" à faire partie de l'échantillon est une notion équivalente à celle de la probabilité de sélection et au "coefficient de propension" (propensity score) de Rosenbaum et Rubin [22]). À partir de cette notion, Rubin [23] a présenté une justification intéressante, dans une optique bayésienne, de l'utilisation de la randomisation dans la sélection d'un échantillon, démarche qui est défendue fermement par les partisans des méthodes basées sur le plan de sondage, mais qui a été accueillie avec une sorte de mépris par beaucoup de défenseurs des méthodes basées sur un modèle. Par conséquent, Rubin préconise l'application de techniques d'inférence basées sur un modèle et une analyse sérieuse des effets du processus de sélection ou de la propension d'inclusion dans l'échantillon; dans certains cas, ces principes conduiraient soit à (2.2), soit à (2.3) ou à d'autres estimateurs possibles.

La deuxième observation concerne le fait que DuMouchel et Duncan ont explicitement limité leur analyse à la question de la pondération dans les échantillons aléatoires simples stratifiés. Un problème qui est aussi important dans un grand nombre de cas est celui de l'effet de grappe, c'est-à-dire la dépendance entre les unités échantillonnées qui est attribuable à leur inclusion simultanée dans l'échantillon en vertu du plan de sondage comme, par exemple, quand des personnes appartenant à un ménage échantillonné ou des personnes appartenant à des ménages voisins sont simultanément choisies pour l'échantillon. Dans les échantillons autopondérés, (dont toutes les unités ont un poids égal), les méthodes basées sur le plan de sondage et celles basées sur un modèle peuvent produire les mêmes estimations pour les paramètres d'un modèle analytique, mais l'évaluation de la fiabilité de ces résultats peut être fort différente, sauf si les dépendances découlant de l'effet de grappe sont explicitement incorporées dans le procédé d'inférence basé sur un modèle. Contrairement au problème de l'utilisation des poids dans les échantillons aléatoires simples stratifiés, où une méthode d'inférence basée sur un modèle peut être justifiée si les termes d'erreur ont exactement

cuté par les logiciens habituels pour la régression linéaire et qui vise à déterminer s'il y a une différence marquée entre les résultats de régressions pondérées et non pondérées. Si ce test nous oblige à rejeter l'hypothèse selon laquelle (2.2) et (2.3) sont des estimateurs convergents d'un même ensemble de coefficients, on doit utiliser l'estimateur (2.3) parce qu'il converge vers la valeur de  $\tilde{\beta}^*$  pour la population. Si cette hypothèse n'est pas rejetée, les auteurs préfèrent (2.2) parce que sa variance est (généralement) plus faible.

Si un chercheur rejette (2.2) à cause des résultats du test proposé par DuMouchel et Duncan et qu'il calcule plutôt (2.3), le motif de ce choix est relativement clair: (2.3) est choisi au lieu de (2.2) parce que (2.3) demeure convergent si le modèle ne peut être accepté. Si ce test permet d'"accepter" l'hypothèse nulle et qu'on utilise (2.2) et les erreurs types correspondantes calculées dans le cadre du modèle, il faut néanmoins être prudent vis-à-vis des interprétations natives faites au sujet du paramètre de la population  $\tilde{\beta}^*$  à partir de (2.2) et des intervalles de confiance construits à l'aide de cet estimateur. Dans bien des cas, le choix de (2.3) et du degré de fiabilité qu'il offre constitue le moyen le plus "sûr" d'interpréter les données et d'estimer  $\tilde{\beta}^*$  quand on pense que le modèle peut être inexact, même si un test de l'hypothèse selon laquelle il n'existe aucune différence significative entre les analyses pondérées et non pondérées ne permet pas de rejeter le modèle.

L'article de DuMouchel et Duncan met clairement en évidence les facteurs les plus essentiels dans le choix entre les techniques d'inférence basées sur un modèle et celles basées sur le plan de sondage, à savoir l'efficacité si le modèle est bien spécifié et la convergence si les hypothèses du modèle sont inacceptables. Il serait opportun de faire deux observations concernant ces méthodes. D'abord, bien que l'omission des poids d'échantillonnage rende les estimateurs non convergents dans tout procédé basé sur le plan de sondage et soit justifiée seulement dans les techniques basées sur un modèle, les techniques basées sur un modèle ne font pas toutes abstraction des informations contenues dans ces poids.

Rubin [24] a formulé une explication concise de ce dernier point dans son célèbre article de Hansen, Madow et Tepping [15]. Faisant référence aux notions plus développées élaborées par Rosenbaum et Rubin [22], Rubin a

où  $\tilde{\varepsilon} = \{\varepsilon_i^T\}$  est un vecteur de résidus  $\varepsilon_i^T \sim N(0, \sigma^2)$  qui sont indépendants et identiquement distribués,  $\tilde{\beta}$  pour estimateur du maximum de vraisemblance de  $\beta$

$$\tilde{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}. \quad (2.2)$$

Dans le calcul des estimations d'une enquête, un poids  $W_i$  est affecté à chaque unité  $i$  de l'échantillon en fonction de l'inverse de la probabilité de sélection, et des facteurs de correction pour les cas de non-réponse et l'estimation par le quotient sont souvent utilisés. Si  $\tilde{W}$  est une matrice diagonale de poids  $W_i$ , l'estimateur

$$\hat{\beta} = (\tilde{X}^T \tilde{W} \tilde{X})^{-1} \tilde{X}^T \tilde{W} \tilde{Y}. \quad (2.3)$$

donne un résultat qui est rectifié pour tenir compte du plan de sondage et des poids d'échantillonnage. Dans le modèle stochastique qui justifie le choix de l'estimateur (2.2) ou, plus généralement, s'il n'existe aucune corrélation entre les valeurs de  $\varepsilon_i^T$  et que ces erreurs ont une espérance mathématique nulle et des variances égales, l'estimateur (2.3) a une plus grande variance d'échantillonnage que (2.2). En revanche, si ces conditions ne sont pas satisfaites (en particulier celle relative à l'espérance mathématique des  $\varepsilon_i^T$ ), l'estimateur (2.3) produit quand même une estimation corrigée en fonction du plan de sondage pour le paramètre de la population  $\tilde{\beta}^*$  et correspond à l'application de (2.2) aux valeurs contenues dans l'ensemble de la population finie, mais le calcul de (2.2) sans pondération des unités de l'échantillon n'assure pas une estimation convergente de  $\tilde{\beta}^*$ .

Dumouchel et Duncan vont plus avant dans leur étude du choix entre les estimateurs (2.2), dont la variance est plus faible dans le modèle simple, et (2.3), qui est convergent même si les hypothèses du modèle s'avèrent fausses. Ils citent d'autres auteurs qui ont pris l'un ou l'autre des deux partis dans cette controverse: leur article est un excellent exposé des différents points de vue sur cette question. Ils proposent également un test qui peut être exé-

Le choix entre une méthode d'inférence basée sur le plan de sondage et une étude basée sur un modèle peut dépendre de plusieurs facteurs tels que les effets de la stratifications et l'existence ou l'ampleur de la dépendance entre les valeurs échantillonnées (effet de grappe). DuMouchel et Duncan [6] ont énuméré les principales questions soulevées par ce choix dans leur étude sur la nécessité d'intégrer les poids d'échantillonnage dans les régressions linéaires.

Si  $\tilde{Y}$  représente un vecteur colonne d'observations  $Y_i$  et que  $\tilde{X} = \{X_{ij}\}$ ,  $j = 1, \dots, p$  représente une série de variables explicatives de  $Y$ , le modèle

## 2. SÉLECTION D'UNE TECHNIQUE D'INFÉRENCE BASÉE SUR LE PLAN DE SONDAJE OU SUR UN MODÈLE

La troisième section décrit brièvement l'expérience du Censur Bureau avec les méthodes de régression linéaire basées sur le plan de sondage. La quatrième section résume un procédé qui est utilisé dans la mise en oeuvre informatique des méthodes de duplication et repose sur des "poids d'échantillonnage artificiels" (replicate weights). Rien que cette technique soit conçue principalement pour calculer la variance des statistiques habituellement estimées dans les enquêtes, elle facilite également le calcul des erreurs types dans les modèles complexes. Cette solution générale peut s'avérer particulièrement utile pour l'utilisation de modèles relativement moins répandus, c'est-à-dire les schémas autres que les modèles linéaires. Les modèles log-linéaires et d'autres modèles linéaires généralisés. Enfin, on résume quelques travaux récents sur les modèles log-linéaires et on mentionne des logiciels conçus pour l'ajustement de ces modèles.



de la plupart des organismes statistiques gouvernementaux comme Statistique Canada et le U.S. Bureau of the Census et de la plupart des grandes maisons de sondage privées. Cette forme d'inférence statistique repose sur le procédé de randomisation utilisé pour prélever un échantillon dans une population finie. Les principes appliqués à la construction d'intervalles de confiance et aux tests d'hypothèses découlent de la théorie des grands échantillons et de la randomisation plutôt que d'un modèle précis. Les manuels classiques, comme ceux de Cochran [4], de Kish [17] et de Hansen, Hurwitz et Madow [14] présentent les éléments de cette théorie. Dans un article récent, Hansen, Madow et Tepping [15] ont décrit comment ces techniques sont supérieures aux méthodes "basées sur un modèle" pour l'inférence à partir de données d'enquête: Särndal [25] et Cassel, Särndal et Wretman [3] ont examiné d'une manière un peu différente le choix entre les méthodes basées sur un modèle et celles basées sur le plan de sondage. Dans l'ensemble, les premiers travaux sur les techniques d'inférence basées sur le plan de sondage visaient l'estimation de totaux, de proportions, de moyennes et de ratios à l'échelle de la population, et la plupart des recherches correspondantes sur les méthodes basées sur un modèle mettent également l'accent sur ces statistiques fondamentales.

Par contre, les schémas analytiques les plus répandus, tels que les modèles de régression linéaire, les modèles log-linéaires et les modèles linéaires généralisés, ont d'abord été formulés en fonction de modèles stochastiques explicites fondés, par exemple, sur la loi normale ou la loi multivariée. Les méthodes d'inférence "classiques" ont progressivement assumé le sens de procédés d'inférence statistique fondés sur une distribution hypothétique (l'adjectif "classique" peut englober le terme "bayésien" dans la présente analyse). Les méthodes d'estimation "robustes" ne comprennent pas de restrictions précises concernant la distribution de la population, mais elles comportent souvent des hypothèses qu'on ne retrouve pas généralement dans les techniques d'enquête. Ainsi, par exemple, les termes d'erreur du modèle sont supposés indépendants et considérés comme provenant d'une population symétrique. Beaucoup de chercheurs qui connaissent l'un ou l'autre de ces modèles analytiques les ont appliqués directement à des données d'enquête sans tenir compte des effets possibles du plan de sondage sur la validité des inférences basées sur les hypothèses distributionnelles courantes. Notre propos ici,

## APPLICATION DE MODÈLES LINÉAIRES ET LOG-LINÉAIRES AUX DONNÉES D'ÉCHANTILLONS COMPLEXES

Robert F. Fay<sup>1</sup>

La plupart des enquêtes menées par des organismes comme Statistique Canada ou le U.S. Bureau of the Census reposent sur des plans de sondage complexes. Les techniques d'inférence statistique basées sur le plan de sondage, qui sont généralement utilisées par ce genre d'organisme pour calculer des moyennes et des totaux, peuvent également s'étendre à l'estimation des paramètres de modèles analytiques. La plus grande partie de cette étude porte sur l'application des méthodes d'inférence basées sur le plan de sondage aux modèles théoriques, mais elle présente également des arguments justifiant le recours à des procédés basés sur le modèle dans certains cas, ce qui explique le fait que ces deux formes d'inférence soient utilisées par l'organisme dont l'auteur fait partie.

Cette étude décrit brièvement l'expérience acquise dans l'extension des techniques d'inférence basées sur le plan de sondage à l'analyse de régression linéaire. Récemment, la méthode des "poids d'échantillonage artificiels" (replicate weighting) a été appliquée à l'estimation de la variance dans diverses enquêtes menées par le Census Bureau. Jusqu'à présent, cette méthode a servi avant tout à calculer la variance de variables statistiques simples, mais elle facilite aussi l'évaluation des variances dans pratiquement n'importe quel modèle analytique complexe. Enfin, on décrit des techniques relatives aux modèles log-linéaires et on résume les travaux faits sur ce sujet.

### 1. INTRODUCTION

Statistique Canada a joué un rôle important dans l'élaboration d'un grand nombre de procédés pour l'application de méthodes analytiques aux données d'enquête. Le présent exposé a pour objet de résumer et de partager l'expérience acquise par le U.S. Bureau of the Census dans ce domaine. Les techniques d'inférence "basées sur le plan de sondage" (aussi appelées méthodes "classiques") prédominent dans l'analyse et la production des données

<sup>1</sup> Robert F. Fay, Statistical Methods Division, U.S. Bureau of the Census, Washington, D.C.



Figure 10: Graphique Diagnostique de  $\{G^2 - G^2(-1)\} / \frac{1}{2}$

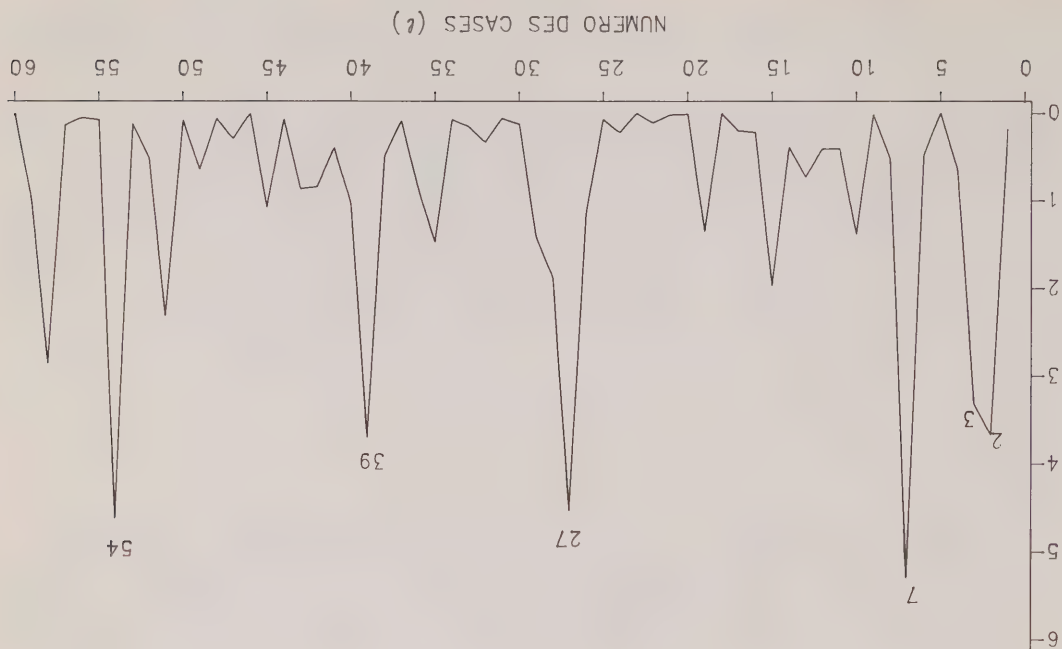


Figure 9: Graphique Diagnostique de  $\{G^2 - G^2(-1)\} / \frac{1}{2}$

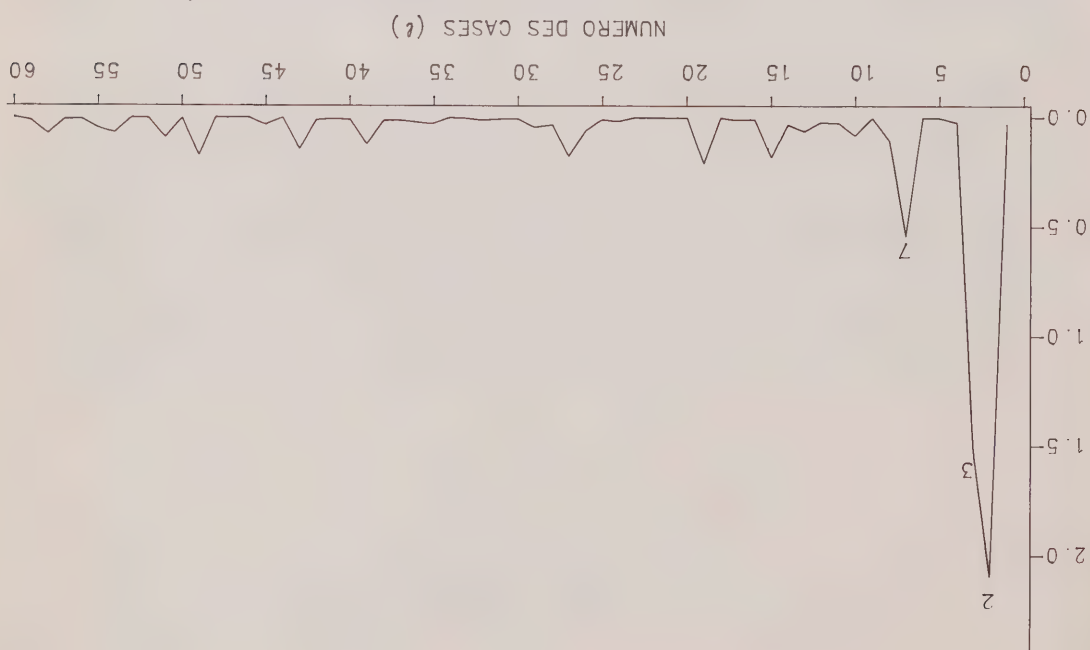


Figure 8: Graphique Diagnostique de  $\{\hat{\beta}_3 - \beta_3(-1)\}/e.t.(\hat{\beta}_3)$

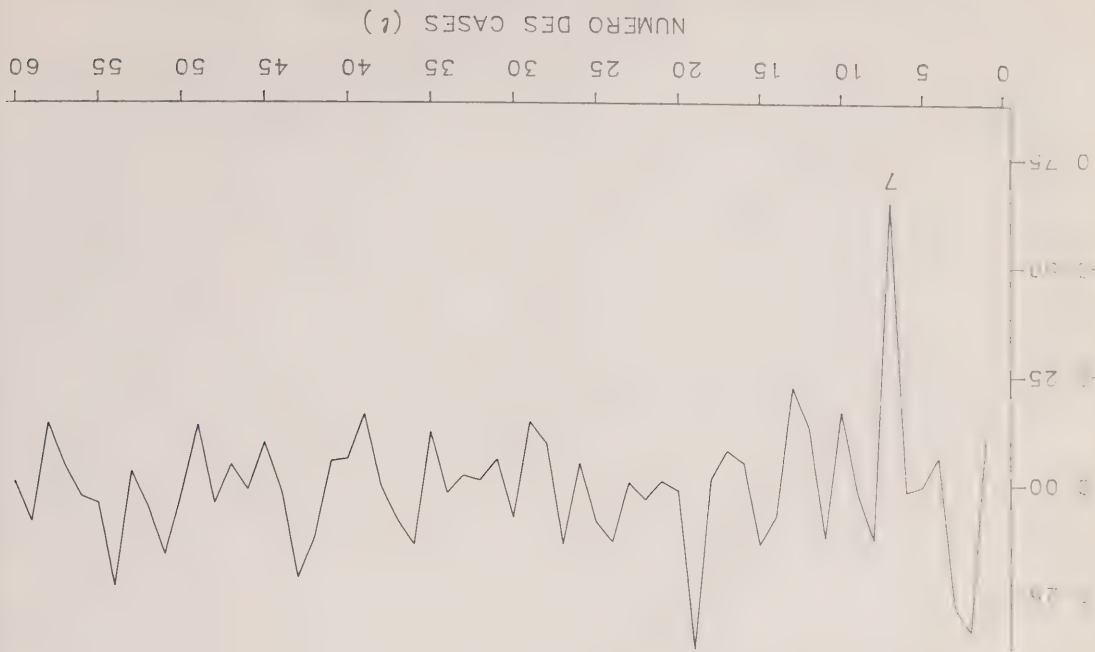
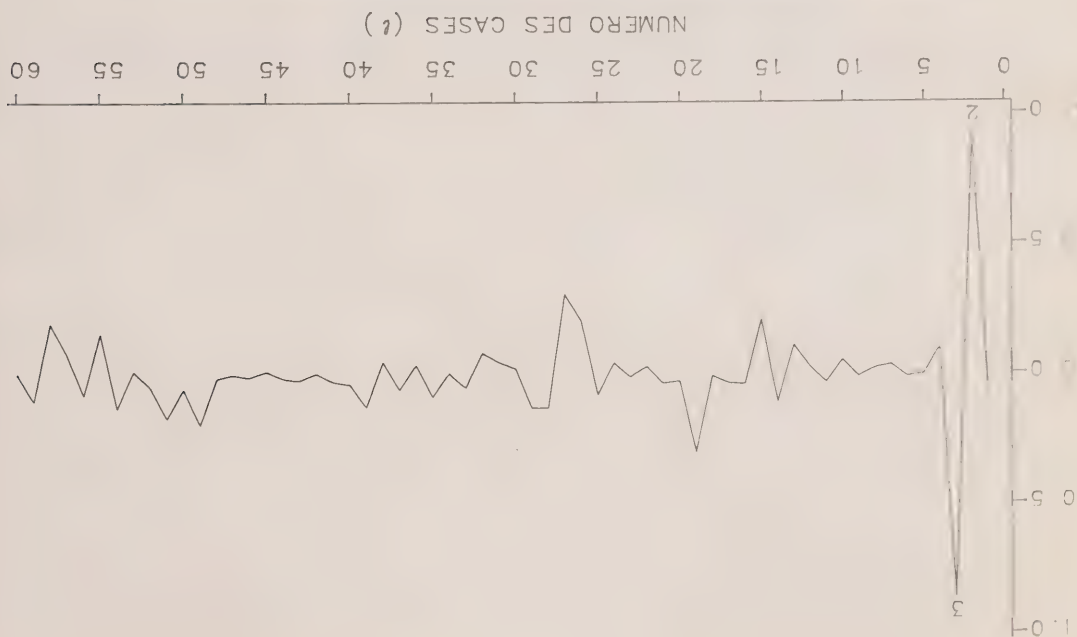
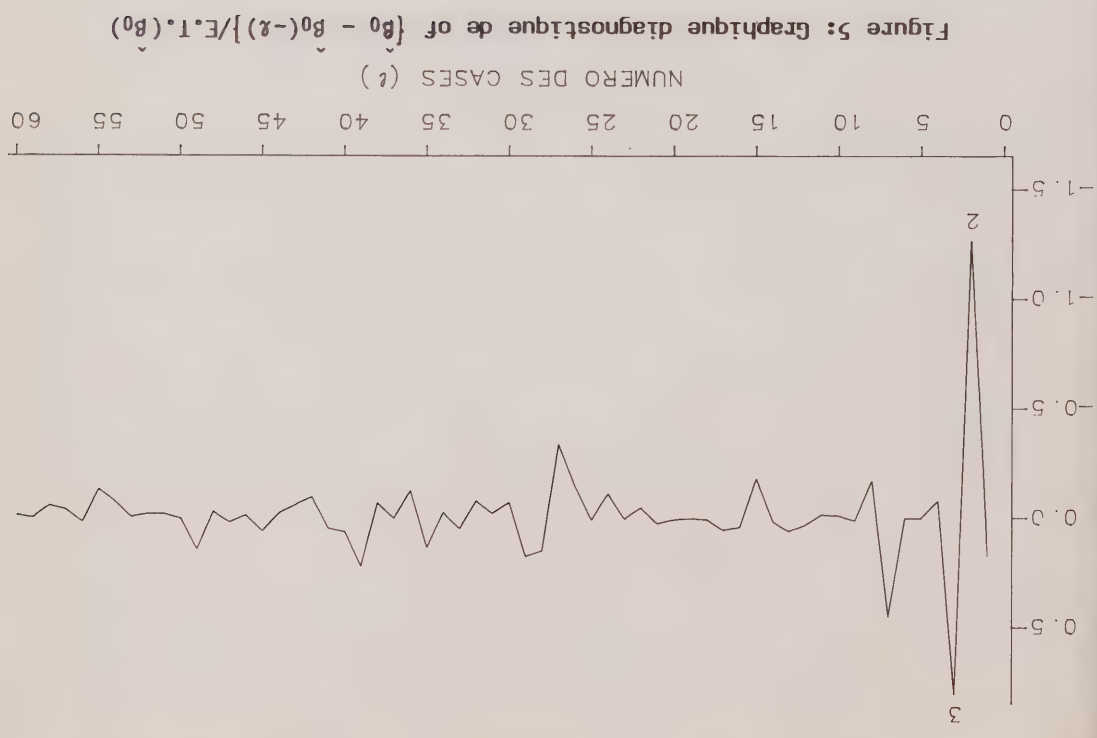
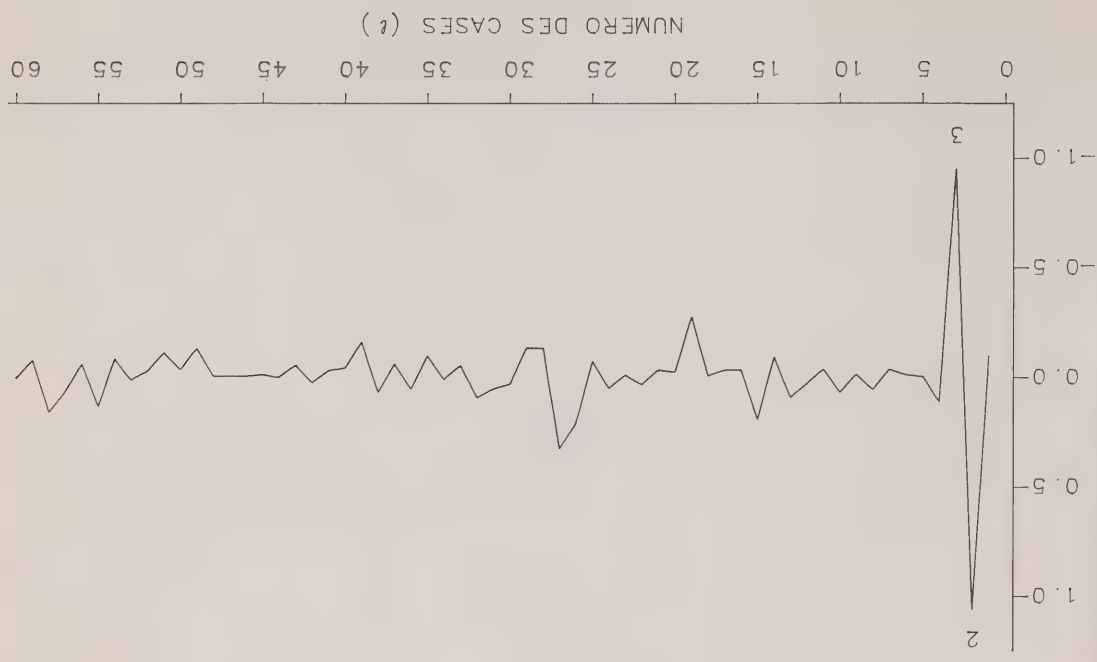


Figure 7: Graphique Diagnostique de  $\{\hat{\beta}_2 - \beta_2(-1)\}/e.t.(\hat{\beta}_2)$





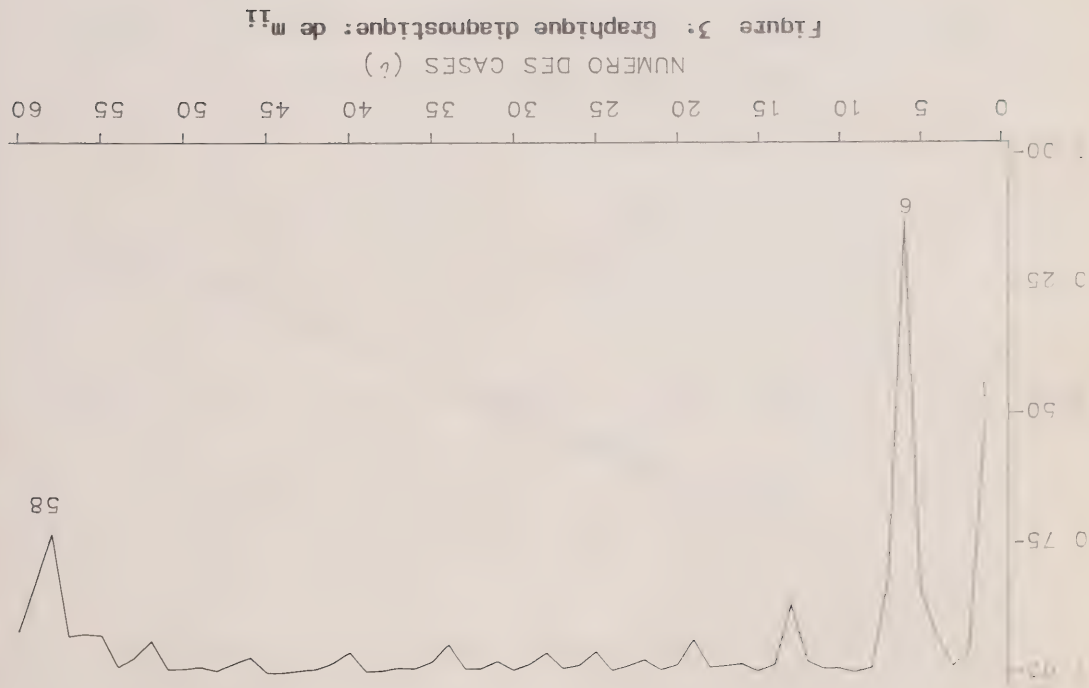


Figure 3: Graphique diagnostique de  $m_{11}$

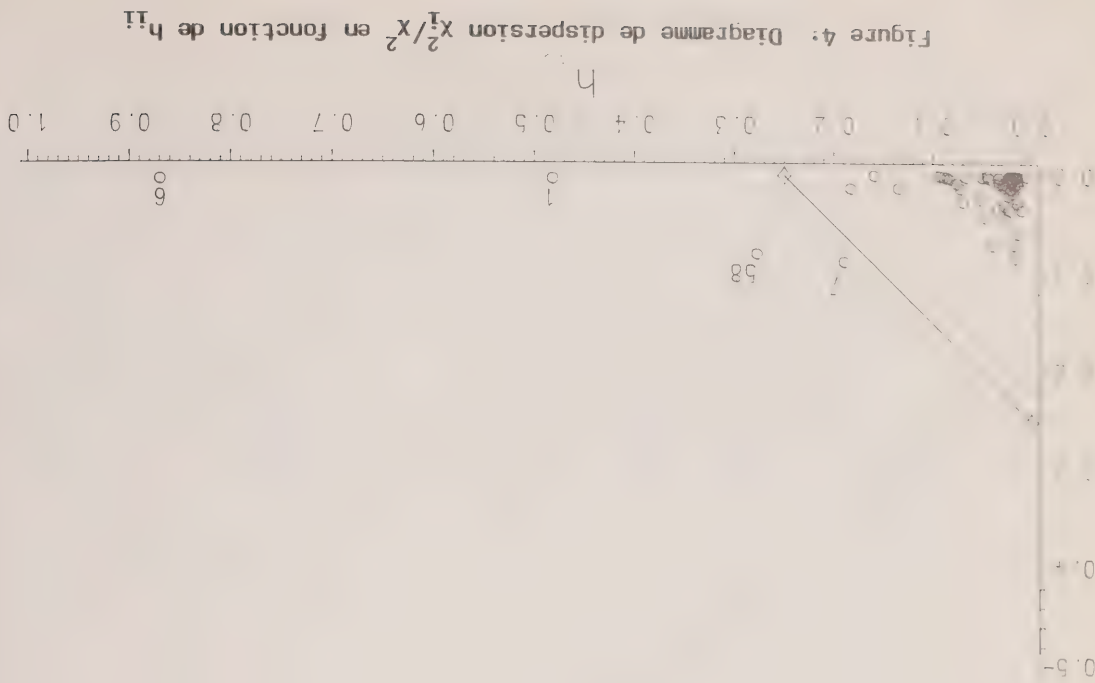
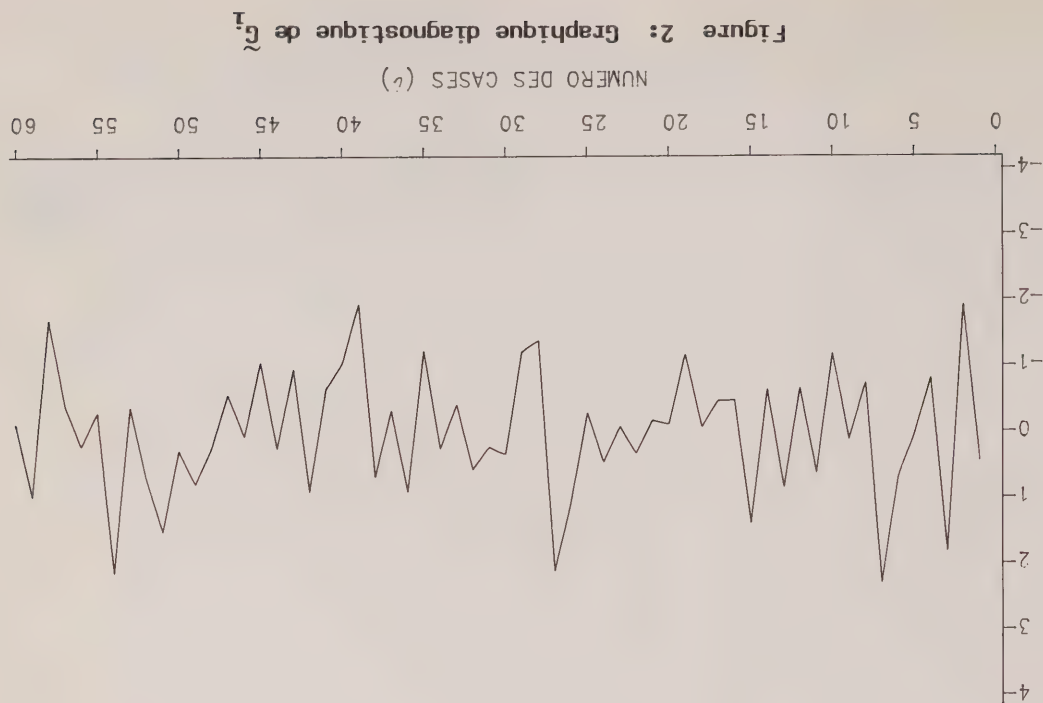
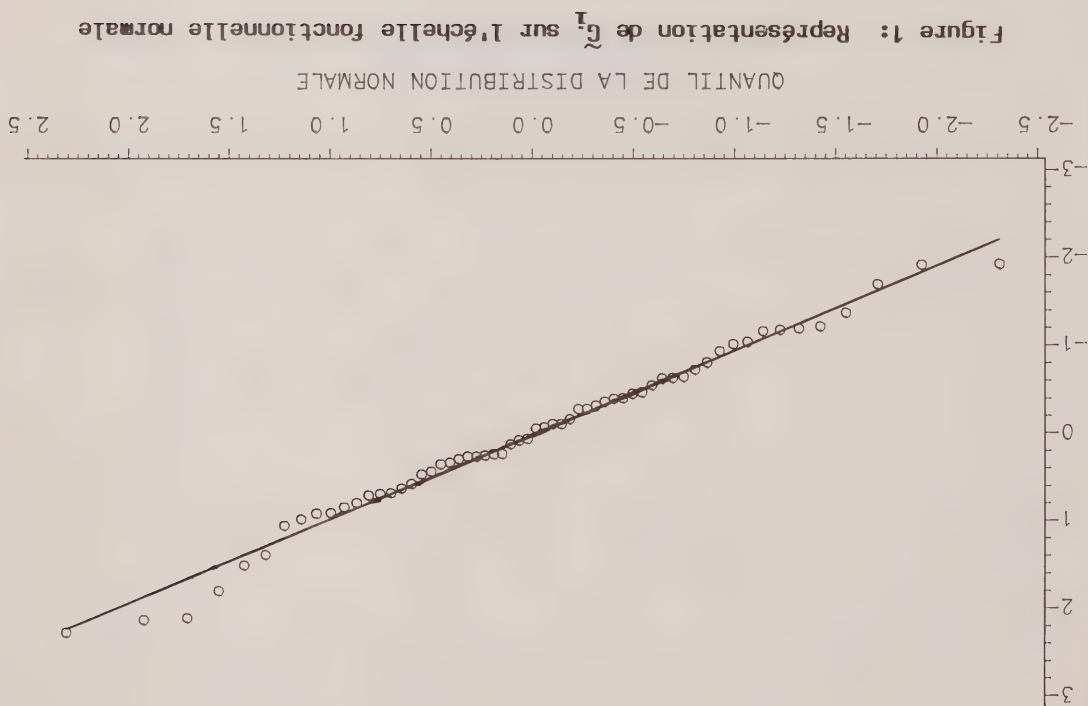


Figure 4: Diagramme de dispersion  $X_1^2/X^2$  en fonction de  $h_{11}$



- [6] Felleq, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. J. American Statistical Association 75, p. 261-268.
- [7] McQuilagh, P. et Neider, J.A. (1983). Generalized Linear Models. Chapman and Hall, Londres.
- [8] Pregibon, D. (1981). Logistics regression diagnostics. Ann. Statist. 9, p. 705-724.
- [9] Rao, J.N.K. et Scott, A.J. (1984). On simple adjustments to chisquared tests with survey data: log-linear and logit models. Manuscrit non publié.
- [10] Roberts, G. (1984). On chi-squared tests for logit models with cell proportions estimated from survey data. Manuscrit non publié. Université Carleton.



1981). La figure 10 révèle des sommets importants pour les cas 2, 3, 7, 27, 39 et 54 (valeurs  $\geq 3$ ), la valeur la plus élevée étant celle de la case 7, qui a un indice égal à 5.4. Si on supprime la case 7 et qu'on calcule de nouveau la variable corrigée  $G_2^c(-x) = G_2(-x)/\delta(-x)$ , où  $\delta(-x)$  est la valeur correspondante de  $\delta$ , on obtient un résultat égal à 48.43 avec 55 degrés de liberté, alors que  $G_2^c/\delta = 55.3$  avec 56 degrés de liberté.

Dans l'ensemble, nos résultats révèlent qu'il serait peut-être opportun de supprimer les cas 7, 2 et 3 mais nous croyons que leur effet sur les estimations n'est pas assez important pour justifier une telle décision. Après tout, on veut expliquer la variation entre les proportions contenues dans toutes les cases, à moins qu'il n'existe un grand désaccord entre les données et le modèle ajusté, qui est attribuable à certaines cases seulement.

## REMERCIEMENTS

Nous aimerions remercier M. Gratton de Statistique Canada qui a fait les graphiques présentés dans cette analyse.

## BIBLIOGRAPHIE

- [1] Bloch, F.E. et Smith, S.P. (1977). Human capital and labour market employment. J. Human Resources 12, p. 550-559.
- [2] Cook, R.D. (1977). Detection of influential observations in linear regression. J. American Statistical Association 72, p. 169-174.
- [3] Cook, R.D. et Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, Londres.
- [4] Cox, D.R. (1970). Analysis of Binary Data. Chapman and Hall, Londres.
- [5] Fay, R.E. (1983). Replication approaches to the log-linear analysis of data from complex samples. Manuscript non publié.

de l'espace et parce qu'elle ressemble à celle de  $\hat{G}_1$ . La figure 1 révèle une allure linéaire dont aucun point ne s'éloigne beaucoup. Le graphique diagnos-  
tique de  $\hat{G}_1$  présenté à la figure 2 concorde avec la figure 1. Il semble donc  
qu'aucun cas ne contient des proportions extrêmes quand on utilise  $\hat{G}_1$  et  $\hat{G}_2$   
pour l'analyse des résidus.

#### (ii) Repérage des cas prépondérantes

Le graphique diagnostique de  $m_{11}$  est présenté à la figure 3, où les points  
des cas 1 et 6 se trouvent bien loin de l'ensemble des points. La figure 4  
montre la relation entre  $X_1^2/X_2^2 = X_1^2/X_2^2$  et  $h_{11}$ , où une droite ayant une pente  
égale à -1 a pour équation  $X_1^2/X_2^2 + h_{11} = 3 \text{ moy}(h_{11}^*)$ . Ici  $h_{11}^* = h_{11} + X_1^2/X_2^2$  et  
les valeurs de  $h_{11}^*$  voisines de un correspondent aux cas qui sont soit extrê-  
mes, soit prépondérantes, voire les deux (Pregibon, 1981) et se trouvent au-  
dessus de la droite tracée à la figure 3. Il est clair que les cas 1 et 6  
et, dans une certaine mesure, les cas 7 et 58 devraient être examinés de  
plus près.

#### (iii) Indice de la sensibilité des coefficients

Les graphiques diagnostiques de l'indice qui mesure la sensibilité des  
coefficients aux différentes cas (relation entre  $\Delta_j(\lambda)$  et  $\lambda$ ) sont présentés  
aux figures 5, 6, 7 et 8 pour  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$  et  $\beta_3$  respectivement. Il ressort  
clairement de ces graphiques que les cas 2 et 3 engendrent de l'instabilité  
dans  $\beta_0$ ,  $\beta_1$  et  $\beta_2$ , tandis  $\beta_3$  est perturbé par la case 7.

#### (iv) Indice de la sensibilité des valeurs ajustées

La figure 9 présente le graphique de la relation entre  $[G^2 - G^2(-\lambda)]/\delta_c = c$   
et  $\lambda$  et permet d'évaluer l'effet des cas individuelles sur les valeurs ajus-  
tées. Cette figure comprend un grand sommet correspondant aux cas 2 et  
3 et un sommet plus bas pour la case 7. Comme Cook (1977) et Pregibon (1981)  
l'ont déjà souligné, on peut comparer  $c$  et le pourcentage lié à  $X^2(s)$  ( $s = 4$   
dans le modèle (28)) pour avoir une idée approximative du contour de la région  
de confiance où la pseudo-FMV est ramenée quand la  $i$ ème case est supprimée.  
Dans ce cas, la valeur 2.1 pour la case 2 correspond à peu près au contour de  
78% de la région de confiance.

#### (v) Indice de la sensibilité de la validité de l'ajustement

La figure 10 montre la relation entre  $[G^2 - G^2(-\lambda)]/\delta_c$  et  $\lambda$ . Le graphique  
de  $[X^2 - X^2(-\lambda)]/\delta_c$  est semblable et, par conséquent, on ne le présente pas  
ici. Le dernier indice est d'ailleurs le moins préférable des deux (Pregibon,

Contrairement à ce qui s'est produit dans le test de la validité de l'ajustement, la variable de Wald est stable dans le test d'hypothèses emboîtées et atteint des valeurs proches des résultats correspondants obtenus pour  $G^2(2|1)/\hat{s}(H)$ .

Le test de la validité de l'ajustement et ces tests d'hypothèses emboîtées permettent de construire un modèle simple comprenant seulement quatre paramètres :

$$v_{jk} = \ln \frac{1 - \pi_{jk}}{\pi_{jk}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k, \quad (28)$$

où  $\hat{\beta}_0 = -3.10$ ,  $\hat{\beta}_1 = 0.211$ ,  $\hat{\beta}_2 = 0.00218$  et  $\hat{\beta}_3 = 0.1509$  et les erreurs types sont 0.247, 0.0130, 0.000172 et 0.0115 respectivement. Nous utilisons le modèle (28) à la section 3.2 pour définir des indices diagnostiques de la régression logistique.

### 3.2 Indices diagnostiques

Nous décrivons maintenant des exemples de l'utilisation des indices diagnostiques décrits à la section 2.4.

#### (1) Analyse des résidus

Les soixante cases du tableau à deux dimensions ont été numérotées en ordre lexicographique et les résidus centrés réduits  $e_i$  ont été calculés pour le modèle (28) qui a été formulé après une série de tests d'hypothèses. Parmi les soixante valeurs de  $e_i$ , celles des cases numéros 6 et 54, où  $p_{jk} = 1$ , étaient très élevées, soit 166.6 et 6.2 respectivement. Parmi les autres  $e_i$ , les résidus numéros 7, 27 et 59 étaient égaux à 3.84, 2.73 et 2.52 respectivement, tandis que l'espérance mathématique du nombre de  $|e_i|$  dépassant 2.35 dans le cadre du modèle (28) est d'environ  $0.02 \times 60 = 1.2$ . Par conséquent, il y a lieu de penser que les cases 7 et 27 contiennent peut-être des proportions extrêmes.

Une représentation graphique de  $\hat{G}_i$  sur une échelle fonctionnelle normale est présentée à la figure 1; la courbe de  $\hat{\chi}_i^2$  n'est pas incluse pour économiser

aléatoire calculées dans le modèle (27), ainsi que la variable de

$H_1: \beta_1 = 0, 1, 2, 3, 4$  à l'intérieur du modèle (27). Tel que prévu, les vraies valeurs des erreurs types sont plus élevées que celles des erreurs types binomiales correspondantes. L'hypothèse  $H_4: \beta_4 = 0$  (autrement dit, l'hypothèse selon laquelle le coefficient de  $E_2^1$  est nul) ne peut être rejeté au seuil de 5% à partir de la variable de Wald ou de la variable  $G_2^2$ . En revanche, le coefficient  $\beta_2$  de  $A_2^1$  est très significatif. Quand on examine le niveau de signification des coefficients individuels, on compare les valeurs de  $\chi^2(2|1)$  ou  $G_2^2(2|1)/\delta.(H)$  et  $\chi^2_{0.05}(1) = 3.84$ , la valeur critique du  $\chi^2$  à un degré de liberté pour laquelle le risque de première espèce est de 5%.

Mais avons aussi vérifié les hypothèses emboîtées suivantes dans le cadre du modèle (27):  $H_3: \beta_3 = \beta_4 = 0$  (ce qui revient à dire que le niveau d'ins-truction n'a aucun effet);  $H_{24}: \beta_2 = \beta_4 = 0$  (hypothèse d'absence d'effets qua-dratiques). Les résultats de  $H_3$  et  $H_{24}$  sont très significatifs:

$$G_2^2(2|1)/\delta.(H_{34}) = 282.2/1.64 = 172.1, \chi^2(2|1) = 165.6 \text{ for } H_{34};$$

$$G_2^2(2|1)/\delta.(H_{24}) = 242.2/2.28 = 106.3, \chi^2(2|1) = 162.1 \text{ pour } H_{24}, \text{ alors que } \chi^2_{0.05}(2) = 5.99.$$

Tableau 1: Pseudo-FMV  $\hat{\beta}_1$ , e.t.  $(\hat{\beta}_1)$ ,  $\chi^2(2|1) = \hat{\beta}_1^2/\text{var}(\hat{\beta}_1)$  et  $G^2(2|1)/\delta.(H_1)$  pour les données de l'EPA à l'intérieur du modèle (27).

$\hat{\beta}_1$	e.t. $(\hat{\beta}_1)$		Valeur Binomiale	Valeur	$\chi^2(2 1)$	$G^2(2 1)/\delta.(H_1)$
	Vraie	Valeur				
0	-2.76	0.557	0.0132	250.6	24.6	168.4
1	0.209	0.0132	0.012	250.6	157.3	102.1
2	-0.00217	0.000173	0.000136	157.3	1.04	1.01
3	0.0913	0.0891	0.068	1.04	0.45	0.46
4	0.00276	0.00411	0.0030	0.45		

Des recherches en sociologie ont aussi révélé que ce genre de modèle est justifié (Bloch et Smith, 1977). À partir des résultats présentés à la section 2, on a obtenu les valeurs suivantes pour les variables utilisées dans le test de la validité de l'ajustement:

$$\begin{aligned} \chi^2 &= 98.9 & G^2/\delta &= 52.5 \\ G^2 &= 101.2 & G^2/\delta &= 53.7, \quad \delta = 1.88. \end{aligned}$$

Étant donné que  $\chi^2$  et  $G^2$  dépassent  $\chi^2_{0.05}(55) = 73.3$ , la valeur critique du  $\chi^2$  à  $I - s = 55$  degrés de liberté pour laquelle le risque de première espèce est de 5%, on rejeterait le modèle si on ne tenait pas compte du plan de sondage. Par contre, la valeur de  $\chi^2/\delta$ , ou  $G^2/\delta$ , indique que le modèle est acceptable à un niveau de signification (seuil  $P$ ) d'approximativement 0.52. La variable  $\chi^2$ , une fois rectifiée en fonction de  $\chi^2_{0.05}(55)$ , est égale à 47.7, valeur non significative. En outre, dans le cas présent, où  $s (= 5)$  est relativement faible en comparaison de  $I (= 60)$ , le facteur de rectification simple  $\bar{d}$ , la moyenne des effets du plan de sondage dans une case (voir l'annexe 1980), est très proche de  $\bar{d} = 1.905$ , tandis que  $\delta = 1.88$  (voir Rao et Scott (1984) pour un exposé théorique).

La variable de Wald  $\chi^2_S$  n'est pas définie dans ce cas parce que  $p_{jk}^1 = 1$  dans deux cases, mais nous avons modifié légèrement les totaux estimés pour nous assurer que  $p_{jk}^1 > 1$  dans toutes les cases et nous avons ensuite calculé  $\chi^2_S$ . Les valeurs de  $\chi^2$  qu'on a obtenues sont toutes grandes en comparaison de  $\chi^2/\delta$ . (elles sont au moins trente fois supérieures à  $\chi^2/\delta$ ) et varient beaucoup (de 1715 à 3061). La variable de Wald est donc très instable dans ce test de la validité de l'ajustement. Si on élimine deux cases où  $p_{jk}^1 = 1$ ,  $\chi^2 = 68.4 > \chi^2_{0.05}(53) = 71.0$ , ce qui indique que le modèle (27) est accepté. Mais supprimer des cases simplement pour justifier l'emploi d'une certaine variable dans un test ne constitue pas une démarche souhaitable. Une autre difficulté liée à  $\chi^2$ , que Fay (1983) a déjà signalée, ne se présente pas dans ce test parce que le nombre de degrés de liberté de  $V$  est grand en comparaison du nombre de cases dans le tableau.

Le tableau 1 présente les valeurs des pseudo-FMV, de leurs erreurs types (e.t.) et des erreurs correspondantes dans l'échantillonnage binomial



groupes d'âge en divisant l'intervalle [15, 64] en dix groupes dont le j<sup>ème</sup> état l'intervalle  $[10 + 5j, 14 + 5j]$ ,  $j = 1, 2, \dots, 10$ , et la valeur au milieu de chaque intervalle,  $A_j$ , a été utilisée comme l'âge de toutes les personnes dans un même groupe d'âge. Pareillement, on a défini des niveaux d'instruction,  $F_k$ , en attribuant à chaque personne un nombre d'années d'instruction correspondant à la médiane dans sa catégorie respective: on a ainsi obtenu six niveaux d'instruction, soit 7, 10, 12, 13, 14 et 16 années. La classification croisée de l'âge et du niveau d'instruction remplit donc un tableau à deux dimensions composé de  $I = 60$  cases contenant chacune une proportion  $\pi_{jk}$ .

L'EPA repose sur un plan d'échantillonnage par drappes stratifié à plusieurs degrés. Deux degrés d'échantillonnage sont définis pour les régions urbaines autorenseignées (AR) et trois ou quatre degrés pour les régions non autorenseignées (NAR), dans chaque province. Les estimations de l'occupation  $\hat{p}_{jk}$  subissent une correction pour tenir compte de la stratification a posteriori en fonction des projections du recensement relatives à la composition par âge et par sexe au niveau provincial. La matrice des covariances estimées  $\hat{V}$  des estimations  $\hat{p}_{jk}$  est calculée à partir de plus de 450 unités au premier degré d'échantillonnage (UPC), de sorte que le nombre de degrés de liberté de  $\hat{V}$  est grand en comparaison de  $I = 60$ .

### 3.1 Tests d'hypothèses

Les graphiques diagnostiques de la relation entre les loqits  $\hat{V}_{jk}$  et l'âge  $A_j$  pour chaque niveau d'instruction  $F_k$  montrent que, pour une valeur donnée de  $k$ ,  $\hat{V}_{jk}$  augmente généralement en fonction de l'âge, atteint un sommet et diminue par la suite (en d'autres mots, les courbes sont convexes et ascendantes jusqu'à un maximum). Par conséquent, le modèle suivant pourrait expliquer la variation des  $\pi_{jk}$ :

$$\pi_{jk} = \ln \frac{1 - \pi_{jk}}{\pi_{jk}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 F_k + \beta_4 F_k^2,$$

$$j = 1, \dots, 10; k = 1, \dots, 6. \quad (27)$$



de l'ajustement tels que les estimations des paramètres,  $\hat{\beta}_j$ , les valeurs ajustées,  $\hat{f}_j$ , et les mesures de la validité de l'ajustement  $\chi^2/\delta$ , ou  $G^2/\delta$ , ou d'autres. Comme Pregibon (1981), nous proposons trois indices qui mesurent l'effet des cas (points) extrêmes sur la qualité de l'ajustement.

1) Indice de la sensibilité des coefficients: Soit  $\hat{\beta}_j(-x)$  la valeur de la pseudo-EMV de  $\beta_j$  calculée sans les données de la  $j^{\text{ème}}$  case. L'indice  $\Delta_j(x) = \hat{\beta}_j - \hat{\beta}_j(-x)/e.t.(\hat{\beta}_j)$  mesure la sensibilité du  $j^{\text{ème}}$  coefficient au  $j^{\text{ème}}$  point. Les graphiques diagnostiques de la relation entre  $\Delta_j(x)$  et  $x$  pour chaque  $j$  sont utiles, mais la quantité de graphiques à examiner peut devenir rébarbative si le nombre de coefficients dans le modèle est élevé.

2) Indice de la sensibilité des valeurs ajustées: Le fait que les estimations des coefficients peuvent changer beaucoup quand on supprime le  $j^{\text{ème}}$  point (ou la  $j^{\text{ème}}$  case) ne signifie pas nécessairement qu'il existe de grands écarts entre les valeurs ajustées  $\hat{f}_j$  et  $\hat{f}_j(-x)$ , le vecteur des valeurs ajustées obtenues sans la  $j^{\text{ème}}$  case; autrement dit,  $\|\hat{f}_j - \hat{f}_j(-x)\|$  peut être faible. On utilise donc  $[G^2 - \tilde{G}^2(-x)]/\delta$ , ou  $[\chi^2 - \tilde{\chi}^2(-x)]/\delta$ , pour évaluer l'effet du  $j^{\text{ème}}$  point sur les valeurs ajustées, où  $\tilde{G}^2(-x)$  et  $\tilde{\chi}^2(-x)$  sont la solution des équations (9) et (8) respectivement quand on remplace  $f_j = f_j(\hat{\beta})$  par  $f_j(-x) = f_j(\hat{\beta}(-x))$ .

3) Indice de la sensibilité de la validité de l'ajustement: La valeur de  $[G^2 - \tilde{G}^2(-x)]/\delta$ , ou  $[\chi^2 - \tilde{\chi}^2(-x)]/\delta$ , où  $G^2(-x)$  et  $\chi^2(-x)$  sont respectivement le rapport de vraisemblance et la variable khi-carré calculés sans la  $j^{\text{ème}}$  case, mesure la sensibilité de la validité de l'ajustement. (Soulignons le fait que  $G^2(-x) \neq \tilde{G}^2(-x)$ .)

### 3. APPLICATION DES RÉSULTATS THÉORIQUES À L'EPA

Nous avons appliqué les techniques décrites plus haut à quelques données de l'enquête sur la population active au Canada (EPA) menée en octobre 1980. L'échantillon était composé d'hommes âgés de 15 à 64 ans qui faisaient partie de la population active et n'étudiaient pas à temps plein. Nous avons choisi deux facteurs, l'âge et le niveau d'instruction, pour expliquer la variation des taux de chômage à l'aide de modèles logit. On a réparti l'échantillon en

Par contre, les résidus centrés réduits  $e_i$  perdent leur fiabilité si  $p_i$  est égal à un ou est proche de un. A l'instar de Freqibon (1981), nous proposons l'utilisation de composantes de  $\chi^2$  ou  $G^2$ , notamment  $\tilde{\chi}_i^2 = \chi_i^2/6\frac{1}{2}$  ou  $\tilde{G}_i^2 = G_i^2/6\frac{1}{2}$ ,  $i = 1, \dots, I$ , pour analyser les résidus en contournant cette difficulté. Dans un cas ou dans l'autre, les valeurs élevées de ces composantes devraient donner une idée d'où se trouvent les cas mal expliquées par le modèle. Il est utile d'examiner ces composantes à l'aide de graphiques diagnostiques (c'est-à-dire des courbes illustrant la relation entre  $\tilde{\chi}_i^2$  et  $i$  et entre  $\tilde{G}_i^2$  et  $i$ ). Les représentations de  $\tilde{\chi}_i^2$  ou  $\tilde{G}_i^2$  sur une échelle gaussienne (qui montrent la répartition de valeurs ordonnées parmi les quantiles de la loi normale centrée réduite) est un autre moyen utile de voir si le modèle est exact (c'est-à-dire si les points forment tous une ligne droite).

Freqibon (1981) a proposé l'utilisation des éléments diagonaux  $m_{ii}$  de la matrice-projection

$$M = I - \frac{1}{N} \sum_{i=1}^N \tilde{V}_i \tilde{V}_i' = I - \frac{1}{N} \sum_{i=1}^N \tilde{V}_i \tilde{V}_i' \tilde{V}_i \tilde{V}_i'$$

= I - H (par exemple)

(26)

pour déceler les points prépondérants, où  $\tilde{V}_i$  est la matrice des covariances estimées dans un échantillon binomial: autrement dit, on peut utiliser  $\text{diag}[p(1-p), \dots, p(1-p)]/(n p_i)$  pour des données d'enquête. La matrice M est normalement nécessaire pour résoudre les équations de vraisemblance (4) par la méthode des moindres carrés itérativement pondérés et les petites valeurs de  $m_{ii}$  indiquent les points extrêmes dans l'espace factoriel. Ici encore, un graphique diagnostique (de  $m_{ii}$  en fonction de  $i$ ) serait utile. Un peut souligner le fait qu'on n'est pas obligé de tenir compte des effets du plan de sondage sur  $m_{ii}$  parce qu'il repose sur une "pseudo-EMV" conçue pour un échantillon binomial. Un autre graphique utile qui résume bien les informations contenues dans les graphiques diagnostiques de  $\tilde{\chi}_i^2$  en fonction de  $i$  et de  $\tilde{G}_i^2$  en fonction de  $i$  est le diagramme de dispersion des coordonnées  $\tilde{\chi}_i^2/\chi^2$  et  $\tilde{G}_i^2/G^2$  où  $h_{ii}$  est le  $i$ ème élément de la diagonale de la matrice (26) (voir Freqibon, 1981).

Les variables diagnostiques  $e_i$ ,  $\tilde{\chi}_i^2$  ou  $\tilde{G}_i^2$  et  $m_{ii}$  permettent de déceler les points extrêmes, mais non d'évaluer l'effet de ces points sur divers aspects

$$A = (X_1' \tilde{A} X_2)'^{-1} [X_1' \tilde{D}(\tilde{w}) \tilde{V} D(\tilde{w}) X_2] (X_1' \tilde{A} X_2)^{-1}. \quad (24)$$

Les résidus centrés réduits  $(\hat{f}_i - \bar{f}_i) / [\hat{V}_{ii}(r)]^{\frac{1}{2}}$  peuvent aussi être calculés. Comme dans le test de la validité de l'ajustement, on peut également produire une meilleure approximation à l'aide de la méthode de Satterthwaite.

Une variable de Wald pour  $H: \beta_2 = 0$  est définie par l'équation

$$X_2' (2|1) = \hat{\beta}_2' [\hat{D}(\hat{\beta}_2)]^{-1} \hat{\beta}_2, \quad (25)$$

où  $\hat{D}(\hat{\beta}_2)$  est la sous-matrice principale de (5) qui correspond à  $\hat{\beta}_2$ . En vertu de l'hypothèse  $H$ ,  $X_2'(2|1)$  suit asymptotiquement une loi de  $\chi^2$  à  $u$  degrés de liberté. En particulier, si  $\beta_2$  est scalaire, on peut considérer  $\hat{\beta}_2/e.t.$  ( $\hat{\beta}_2$ ) comme une variable de type  $N(0,1)$  si l'hypothèse  $H: \beta_2 = 0$  est vraie ou  $\hat{\beta}_2'/\text{Var}(\hat{\beta}_2)$  comme une variable  $\chi^2$  à un degré de liberté.

## 2.4 Indices diagnostiques

Il est souhaitable de faire une évaluation critique de l'ajustement du modèle logit en repérant les cas contenant des proportions extrêmes et les points prépondérants dans l'espace factoriel. Pour ce faire, le vecteur des résidus et une matrice-projection de l'espace factoriel sont des ressources utiles. Toutefois, contrairement au modèle linéaire classique, le choix duit des résidus qui peuvent être définis sur différentes échelles. Le choix naturel qui permet de prendre en compte le plan de sondage est le vecteur des résidus centrés réduits  $e_i = (f_i - \bar{f}_i) / [\hat{V}_{ii}(r)]^{\frac{1}{2}}$  présenté à la section 2.1. Étant donné que les  $e_i$  ont approximativement une distribution de type  $N(0,1)$  dans le modèle (1), le nombre de résidus  $e_i$  dont la valeur absolue dépasse 1.96, 2.33 et 2.58 a pour espérance mathématique 0.051, 0.021 et 0.011 respectivement, où  $I$  est le nombre de résidus (cas). Ces probabilités théoriques offrent des critères généraux pour déceler les valeurs extrêmes. Si on fait abstraction du plan de sondage et qu'on utilise les résidus centrés réduits dans un échantillon binomial, on risque de tirer des conclusions erronées.

$$G^2(2|1) = 2n \sum_{i=1}^I w_i \left\{ \frac{f_i}{\hat{f}_i} \ln \frac{f_i}{\hat{f}_i} + (1 - \frac{f_i}{\hat{f}_i}) \ln \frac{(1 - \frac{f_i}{\hat{f}_i})}{(1 - \frac{\hat{f}_i}{\hat{f}_i})} \right\} \quad (19)$$

... vivent. Dans l'échantillonnage hiérarchique,  $\chi^2(2|1)$  et  $G^2(2|1)$  suivent approximativement une loi de  $\chi^2$  à u degrés de liberté quand l'hypothèse H est vraie, mais, dans les plans de sondage généraux, cette propriété n'existe pas. En fait,  $\chi^2(2|1)$  ou  $G^2(2|1)$  a asymptotiquement la même distribution que la somme pondérée,  $\sum \delta_i(H) Z_i$ , de variables indépendantes  $Z_i$  de type  $\chi^2$ , où les poids  $\delta_i(H)$  ( $i = 1, \dots, u$ ) sont les valeurs propres de la matrice des effets du plan de sondage.

$$(\tilde{X}_1' \Delta \tilde{X}_2)^{-1} (\tilde{X}_1' D(\tilde{W}) V D(\tilde{W}) \tilde{X}_2), \quad (20)$$

où

$$\tilde{X}_2 = [I - X_1(X_1' \Delta X_1)^{-1} X_1' \Delta] X_2, \quad (21)$$

(Roberts, 1984). Dans un échantillon hiérarchique, la matrice des effets du plan (20) se ramène à I, comme dans le test de la validité de l'ajustement à la section précédente.

On peut obtenir une simple modification de  $\chi^2(2|1)$  ou  $G^2(2|1)$  en considérant  $\chi^2_c(2|1) = \chi^2(2|1)/\delta \cdot (H)$  ou  $G^2_c(2|1)/\delta \cdot (H)$  comme suivant une loi de  $\chi^2$  à u degrés de liberté si l'hypothèse H est vraie, où

$$u \cdot \delta \cdot (H) = n \sum_{i=1}^I \tilde{V}_{ii}(\tilde{r}) w_i / \hat{f}_i (1 - \frac{\hat{f}_i}{\hat{f}_i}) \quad (22)$$

est le i-ème élément de la diagonale de la matrice des covariances des résidus,  $r_i(H) = \hat{f}_i - f_i$ , définie par l'expression

$$\tilde{V}(r) = D(\tilde{f}) D(1 - \tilde{f}) \tilde{X}_2 \tilde{X}_2' D(\tilde{f}) D(1 - \tilde{f}), \quad (23)$$

où  $\hat{V}_{ij}(r)$  est le  $(i, j)$  ième élément de  $\hat{D}(r)$ . Les variables  $\chi^2_S$  et  $G^2_S$  permettent une correction pour tenir compte de la variation des  $\delta_i$ . Une variable de Wald pour vérifier la validité de l'ajustement du modèle (1) est

$$\chi^2_2 = \hat{V}' G \hat{V} \phi' G' \hat{V} \quad (15)$$

où  $\hat{V}$  est le vecteur de logs  $\hat{V}_i = \log \hat{p}_i$ . La variable  $\chi^2_2$  suit une loi de  $\chi^2$  à 1 - s degrés de liberté dans les grands échantillons. La variable  $\chi^2_2$  n'est pas définie si  $\hat{p}_i = 0$  ou 1 pour une valeur de  $i$ . En outre, cette fonction devient instable lorsqu'une valeur de  $\hat{p}_i$  est proche de un (voir la section 3) ou quand le nombre de degrés de liberté de  $V$  n'est pas grand en comparaison de  $1 - s$  (Fay, 1983).

## 2.3 Hypothèses emboîtées

Supposons que la matrice  $X$  comprend deux partitions  $(X_1, X_2)$ , où  $X_1$  est de dimension  $1 \times r$  et  $X_2$  de dimension  $1 \times u$  (r + u = s); le modèle (1) peut s'écrire

$$\tilde{y} = X_1 \tilde{\beta}_1 + X_2 \tilde{\beta}_2, \quad (16)$$

où  $\tilde{\beta}_1$  est de dimension  $r \times 1$  et  $\tilde{\beta}_2$  de dimension  $u \times 1$ . Nous voulons souvent vérifier l'hypothèse nulle  $H: \tilde{\beta}_2 = 0$  dans le modèle (16). Les pseudo-FMV pour H provisionnement du système d'équations

$$\chi^2_1 D(w) \tilde{f} = \chi^2_1 D(w) \tilde{p} \quad (17)$$

où  $\tilde{f} = f(\tilde{\beta})$  et on obtient encore une fois les solutions par une série de calculs itératifs. Les variables habituelles des tests du khi-carré et du rapport de vraisemblance de  $H: \tilde{\beta}_2 = 0$  sont

$$\chi^2_2(2|1) = n \sum_{i=1}^I \frac{w_i (f_i - \tilde{f}_i)^2}{\tilde{f}_i (1 - \tilde{f}_i)} \quad (18)$$



de  $\chi^2$  à  $I - s$  degrés de liberté, mais pour les plans de sondage généraux, cette propriété n'est pas valable. En fait,  $\chi^2$  (ou  $G^2$ ) a la distribution asymptotique d'une somme pondérée,  $\sum \delta_i^2 Z_i$ , de variables  $\chi^2$ ,  $Z_i$ , ayant chacune un degré de liberté, où les poids  $\delta_i^2$  ( $i = 1, \dots, I - s$ ) sont les valeurs propres d'une matrice des "effets généralisés du plan de sondage" calculée à partir de l'expression  $\Sigma_0^{-1} \Sigma \phi$ , où

$$\Sigma_1 = G[D(\tilde{F})^{-1}D(1 - \tilde{F})^{-1}D(\tilde{F})^{-1}D(1 - \tilde{F})^{-1}]^{-1}G, \quad (10)$$

$$\Sigma_0 = \frac{1}{n} G' A^{-1} G \quad (11)$$

et  $G$  est une matrice quelconque  $I \times (I - s)$  de rang  $I - s$  pour laquelle  $G'X = 0$ ; autrement dit, la matrice  $G$  est orthogonale par rapport à  $X$ . Dans un échantillon binomial,  $\Sigma_0^{-1} \Sigma \phi$  se ramène à  $I$ , la matrice unité. Il peut apporter une modification simple à  $\chi^2$  ou  $G^2$  (Roberts, 1984) en considérant  $\chi^2 = \chi^2/\delta$ , ou  $G^2 = G^2/\delta$ , comme une variable  $\chi^2$  à  $I - s$  degrés de liberté si le modèle est vrai, où

$$(I - s)\delta = n \sum_{i=1}^I \hat{V}_{ii}(x) w_i / [\hat{f}_i(1 - \hat{f}_i)]. \quad (12)$$

La variable corrigée  $\chi^2$  ou  $G^2$  devrait être acceptable, sauf dans les cas où  $\delta$  est un grand coefficient de variation (CV). Dans une meilleure solution, qui repose sur l'approximation de Satterthwaite, la variable  $\chi^2$  à  $I - s$  degrés de liberté, ou

$$a^2 = \Sigma (\delta_i^2 - \delta^2) / [(I - s)\delta^2] \quad (13)$$

est le carré du CV des  $\delta_i^2$  et

$$\sum_{i=1}^I \delta_i^2 = \sum_{i=1}^I \hat{V}_{ii}(x) w_i / [\hat{f}_i(1 - \hat{f}_i)] \quad (14)$$



dans les grands échantillons, où  $\hat{\Delta} = \text{diag} (w_1 \hat{f}_1(1 - \hat{f}_1), \dots, w_I \hat{f}_I(1 - \hat{f}_I))$ . Les éléments diagonaux de la matrice définie en (5) sont les estimations de la variance des estimations  $\hat{\beta}_I$ . De même, la matrice des covariances estimées du vecteur des résidus  $\tilde{r} = \tilde{p} - \tilde{f}$  est

$$\hat{D}(\tilde{r}) = A \hat{V} A' \quad (6)$$

où

$$A = I - D(\tilde{f})D(1 - \tilde{f})X(X'\hat{\Delta}X)^{-1}X'D(\tilde{w}) \quad (7)$$

Les éléments diagonaux  $\hat{V}_{II}(\tilde{r})$  de la matrice définie en (6) permettent de calculer les résidus centrés réduits  $r_I/e \cdot t \cdot (r_I)$ , qui sont utiles pour détecter les cas contenant des proportions extrêmes.

## 2.2 Tests de la validité de l'ajustement

La variable khi-carré normalement utilisée pour vérifier la validité de l'ajustement du modèle (1) est

$$\chi^2 = n \sum_{i=1}^I \frac{(p_i - \hat{f}_i)^2 w_i}{\hat{f}_i(1 - \hat{f}_i)} = \sum_{i=1}^I \chi_{i,2}^2 \quad (8)$$

Le test du rapport de vraisemblance repose sur la fonction

$$G^2 = 2n \sum_{i=1}^I w_i \{ p_i \ln \frac{p_i}{\hat{f}_i} + (1 - p_i) \ln \frac{(1 - p_i)}{(1 - \hat{f}_i)} \} = \sum_{i=1}^I G_{i,1}^2 \quad (9)$$

Souignons le fait que la variable  $G_{i,1}^2$  est définie quand  $p_i = 0$  et  $1$  et vaut  $-2n \ln(1 - \hat{f}_i)$  et  $-2n \ln \hat{f}_i$  respectivement. Dans l'échantillonnage binomial indépendant, il est bien connu que  $\chi^2$  et  $G^2$  suivent asymptotiquement une loi

viennent des solutions aux équations de vraisemblance suivantes :

$$X'D(\tilde{n}/n)\tilde{f} = X'D(\tilde{n}/n)\tilde{q}, \quad (2)$$

où  $X' = (\tilde{x}_1, \dots, \tilde{x}_I)$ ,  $D(\tilde{n}/n) = \text{diag} (n_1/n, \dots, n_I/n)$ ,  $\tilde{f} = \tilde{f}(\tilde{\beta}) = (f_1, \dots, f_I)'$ , et  $\tilde{q}$  est le vecteur des proportions de l'échantillon tiré du  $i^{\text{ème}}$  domaine ( $\sum n_i = n$ ). Pour les  $i$  de sondage généraux, il n'y a pas d'EMV à cause de difficultés relatives à la formulation de fonctions de vraisemblance convenables. On utilise souvent des "pseudo-EMV" de  $\tilde{\beta}$  ou de  $\tilde{f}$  en remplaçant  $n_i/n$  dans l'équation (2) par l'estimation de la taille relative du domaine,  $w_i = n_i/\tilde{N}$ , et  $q_i$  par l'estimation  $p_i$  calculée à partir des données d'enquête :

$$X'D(\tilde{w})\tilde{f} = X'D(\tilde{w})\tilde{p}. \quad (3)$$

Les estimations qui proviennent de cette équation,  $\hat{\beta}$  et  $\hat{f} = \tilde{f}(\hat{\beta})$ , sont généralement (c'est-à-dire dans de grands échantillons) convergentes. On peut également récrire l'équation (3) sous la forme suivante :

$$X'\tilde{N}_1(m) = X'\tilde{N}_1, \quad (4)$$

où  $\tilde{N}_1$  est le vecteur des totaux estimés  $\tilde{N}_1$  et  $\tilde{N}_1(m)$  est le vecteur de pseudo-EMV,  $\tilde{N}_1^{f_I}(m) = \tilde{N}_1^{f_I}$ , des totaux  $\tilde{N}_1^{f_I}$ . On obtient les valeurs de l'estimation  $\hat{\beta}$ , et, par conséquent, celles de  $\tilde{f}$  et de  $\tilde{N}_1(m)$ , au moyen de l'équation (3) ou (4) par une série de calculs itératifs.

## 2.1 Estimation des variances et des covariances

Soit  $V$  la matrice des covariances estimées de  $\tilde{p}$ ; la matrice des covariances estimées de  $\tilde{\beta}$  est

$$D(\hat{\beta}) = (X'\tilde{\Delta}X)^{-1}(X'\tilde{D}(\tilde{w})V D(\tilde{w})X)(X'\tilde{\Delta}X)^{-1} \quad (5)$$

nous recourons à deux corrections simples de  $\chi^2$  et de  $G^2$  qui reposent sur certains effets généralisés du plan de sondage. Pour analyser des données de l'enquête sur la population active au Canada (EPA) menée en octobre 1980 (section 3). On examine également la variable de Wald qui renferme aussi une correction pour tenir compte du plan de sondage.

En plus de formuler des tests statistiques généraux, il est essentiel de préciser des mécanismes diagnostiques pour déceler les proportions extrêmes dans les cases d'un tableau et les points prépondérants dans l'espace factorel. La recherche de variables diagnostiques pour le modèle linéaire classique a été décrite en détail dans les ouvrages statistiques (voir le livre récent de Cook et Weisberg (1982)). Pregibon (1981) a proposé dernièrement des méthodes semblables pour la régression logistique avec des proportions binomiales. À la section 4, quelques-unes de ces techniques sont appliquées aux données de l'EPA d'octobre 1980 une fois effectuées les corrections nécessaires pour tenir compte du plan de sondage.

## 2. RÉSULTATS THÉORIQUES

Supposons que la population mère est répartie en  $I$  cases (domaines) en fonction du niveau d'un facteur ou plus et que  $N_i$  est l'estimation de la taille du  $i$ ème domaine  $N_i(i = 1, 2, \dots, I; \Sigma N_i = N)$  évaluée à l'aide des données d'une enquête. L'estimation correspondante du total d'une variable binaire  $N_{i1}$  (0, 1) pour le  $i$ ème domaine est représentée par  $N_{i1}$ . Le quotient  $p_i = N_{i1}/N_i$  sert à estimer la proportion  $\pi_i = N_{i1}/N_i$  à l'échelle de la population.

Un modèle logit des proportions  $\pi_i$  a la forme  $\pi_i = f_i(\beta)$ , où

$$\ln \{f_i / (1 - f_i)\} = \text{logit } f_i = \sum_{i=1}^I \tilde{x}_i \beta_i, \quad i = 1, \dots, I. \quad (1)$$

Dans l'équation (1),  $\tilde{x}_i$  est un vecteur de dimension  $s$  de constantes connues calculées à partir des niveaux des facteurs et  $\beta$  est un vecteur de dimension  $s$  de paramètres inconnus. Si un échantillon binomial indépendant est prélevé dans chaque domaine, les estimations du maximum de vraisemblance (EMV) pro-

## RÉGRESSION LOGISTIQUE ET ANALYSE DE DONNÉES DE L'ENQUÊTE SUR LA POPULATION ACTIVE

S. Kumar et J.N.K. Rao<sup>1</sup>

Les tests habituels du khi-carré ( $\chi^2$ ) et du rapport de vraisemblance ( $G^2$ ) pour l'analyse de régressions logistiques comportant une variable binaire sont rectifiés en fonction du plan de sondage. Ces corrections reposent sur certains effets généralisés du plan de sondage. Les variables modifiées sont utilisées pour analyser des données de l'enquête sur la population active au Canada (EPA) menée en octobre 1980. La variable de Wald, qui permet aussi de prendre en compte le plan de sondage, est également examinée pour vérifier la qualité de l'ajustement du modèle et des hypothèses concernant les paramètres du modèle théorique. À l'aide de données de l'EPA, des indices diagnostiques des régressions logistiques sont appliqués à la recherche des proportions extrêmes dans les cases d'un tableau et des points prépondérants dans l'espace factoriel, une fois effectuées les corrections nécessaires pour tenir compte du plan de sondage.

### 1. INTRODUCTION

Les spécialistes des sciences sociales, des sciences du comportement et des sciences de la santé utilisent beaucoup les modèles de régression logistique pour étudier la variation de proportions binomiales (voir, par exemple, les ouvrages de Cox (1970) et McCullagh et Nelder (1983)). Mais quand un plan de sondage comprend des grappes et des strates, les méthodes statistiques appliquées aux proportions binomiales sont souvent inadéquates pour l'analyse des données d'enquête. Par exemple, les tests habituels du khi-carré ( $\chi^2$ ) ou du rapport de vraisemblance ( $G^2$ ) font beaucoup grossir le risque de première espèce (seuil de signification). Il est donc nécessaire de corriger les méthodes classiques en fonction du plan de sondage si on veut faire des inférences valables à partir des données d'une enquête. Dans le présent exposé,

<sup>1</sup> S. Kumar, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada et J.N.K. Rao, département des mathématiques et de la statistique, Université Carleton.

Tableau 6: Résultats de l'analyse de régression  
Modèle 3 - Dépenses par pièce par degré-jour

	1974	1978
Coordonnée à l'origine	.017 A	.02 A
Dépenses totales	8.01×10 <sup>-8</sup> B	1.4×10 <sup>-7</sup> A
Âge du chef de famille	1.8×10 <sup>-5</sup>	9.9×10 <sup>-5</sup> A
Taille de la famille	-1.4×10 <sup>-5</sup>	27.0×10 <sup>-5</sup>
Instruction du chef de famille	-5.3×10 <sup>-4</sup> A	-4.7×10 <sup>-4</sup> A
Type de logement -		
maisons jumelées	3.4×10 <sup>-4</sup>	-7.5×10 <sup>-4</sup>
maisons en rangée	-23×10 <sup>-4</sup> C	-35.9×10 <sup>-4</sup> A
duplex	16.9×10 <sup>-4</sup>	6.3×10 <sup>-4</sup>
R <sup>2</sup> ajusté =	.01	.02
F =	5.6(.0001)	6.6(.0001)
		0.03
		9.5(.0001)

Notez: A = probabilité inférieure à 0.0001, B = probabilité inférieure à 0.001  
C = probabilité inférieure à 0.01

Tableau 5: Résultats de l'analyse de régression  
Modèle 2 - Dépenses par pièce

	1974	1974	1978
	Non pondéré	Pondéré	Non pondéré
Coordonnée à l'origine	76.2 A	77.3 A	99.8 A
ville - Saint-Jean (I.-N.)	30.4 A	32.0 A	74.8 A
Halifax	16.8 A	16.3 B	31.6 A
Montréal	4.5	6.7	6.5
Toronto	-3.5	-1.7	10.1
Winnipeg	-17.6 A	-16.3 A	-0.9
Montreal	-37.9 A	-36.8 A	-26.4 A
Vancouver	0.3	0.8	-6.7
Dépenses totales	$2.2 \times 10^{-4}$ B	$2.5 \times 10^{-4}$ B	$6.9 \times 10^{-4}$ A
Age du chef de famille	0.015	-0.03	0.33 B
Taille de la famille	0.6	0.04	-0.63
Instruction du chef de famille	-1.9 A	-1.4 B	-4.0 A
Type de logement -			
maisons jumelées	-6.5 C	-7.1 B	3.1
maisons en rangée	-11.5 A	-11.8 A	-11.0
duplex	6.1 C	6.6 C	3.24
R <sup>2</sup> ajusté =	.31	.19	.24
F =	73.85(.0001)	38.9(.0001)	41.4(.0001)

Limites de probabilité inférieure à 0.0001, B = probabilité inférieure à 0.001, C = probabilité inférieure à 0.01



Tableau 4: Résultats de l'analyse de régression  
Modèle 1 - Dépenses consacrées à l'énergie à la maison.

	1974	1978
Non pondéré	Pondéré	Non pondéré
Coordonnée à l'origine	197.3 A	225.4 A
Nombre de pièces	13.9 A	12.0 A
Ville - Saint-Jean (T.-N.)	193.9 A	204.9 A
Halifax	75.5 A	73.9 B
Montréal	12.2	22.7
Toronto	-10.2	-3.0
Winnipeg	-127.1 A	-125.4 A
Edmonton	-244.9 A	-243.2 A
Vancouver	-22.9	-17.5
Dépenses totales	.006 A	.006 A
Âge du chef de famille	1.2 A	0.8 B
Taille de la famille	13.2 A	12.1
Instruction du chef de famille	0.7	0.6
Type de logement -		
maisons jumelées	-50.9 B	-49.0 A
maisons en rangée	-81.2 A	-88.9 A
duplex	-12.3	-13.7
R <sup>2</sup> ajusté =	0.43	0.34
F =	118.5(.0001)	79.7(.0001)
		0.38
		74.6(.0001)

Notez: A = probabilité inférieure à 0.0001, B = probabilité inférieure à 0.01  
0.001, C = probabilité inférieure à 0.01

Tableau 2: Rang parmi les groupes types

	Souris		Hippopotames		Lièvres	
	1974	1978	1974	1978	1974	1978
Instruction du chef (faible à élevée)	1	1	9	7	7	8
Age (âge à jeune)	1	2	6	6	9	9
Maris à plein temps (faible à élevé)	1	1	8.5	9	7	6.5
Taille de la famille (faible à élevée)	1	1	9	9	4	4
Enfants de moins de cinq ans (faible à élevé)	3	1	4	2	1.5	7
Enfants de cinq à quinze ans (faible à élevé)	1	2.5	7	6.5	5	2.5
Aliments achetés à des magasins (faible à élevé)	1	1	9	9	4	4
Aliments consommés à des restaurants, etc. (faible à élevé)	1	1	9	9	6	6
Logement (faible à élevé)	1	1	9	7	4	3
Habillage (faible à élevé)	1	1	9	9	6	5
Soins personnels (faible à élevé)	1	1	9	9	5	4
Frais médicaux (faible à élevé)	1	1	8	8	4	4
Tabac et alcools (faible à élevé)	1	1	9	9	7	4
Lecture, loisirs, éducation (faible à élevé)	1	1	9	8	8	9

Tableau 3: Dépenses moyennes consacrées à l'énergie à la maison

	1974	1978	Changement en %
Dépenses moyennes (\$) consacrées à l'énergie de la maison	451	764	+69
Dépenses moyennes par pièce (\$) consacrées à l'énergie à la maison	73	121	+66
Dépenses moyennes par pièce et degré -jours consacrées à l'énergie à la maison	.019	.029	+53

Tableau 1: Groupes types de consommateurs d'énergie

Consommation d'énergie à la maison		Faible 127 mil. kJ	Moyenne 127-222 mil. kJ	Élevée 222 mil. kJ	Total
Consom- mation d'essence pour auto- mobiles	Faible 1136 litres	SOURIS D'ÉGLISE 4.5% de l'échantillon	9.8% de l'échantillon	OURS 2.5% de l'échantillon	16.8
	Moyenne 1136-4545 litres	14.5% de l'échantillon	CASTOR 33.7% de l'échantillon	12.3% de l'échantillon	60.5
	Élevée 4546 litres	LITVRE 4.0% de l'échantillon	12.6% de l'échantillon	HIPPOTAME 6.1% de l'échantillon	22.7
	Total	23.0	56.1	20.9	100.0

Source : Voir la référence à la fin du texte.

comme l'instruction du chef de famille et la variable maison en rangée. Un fait important à noter est la hausse de la valeur du coefficient  $R^2$  ajusté. En fait, les variables indépendantes qui demeurent dans l'équation ne contribuent pas beaucoup à expliquer la variation de la variable dépendante. Il faudrait chercher d'autres variables plus utiles.

En comparant les résultats non pondérés de 1974 et de 1978, dans le cas du modèle 1, on note un changement au niveau du paramètre Vancouver, ainsi qu'une variation de l'importance des maisons jumelées et des duplex par rapport aux maisons individuelles.

Dans le cas du modèle 2, le changement principal réside encore une fois dans l'effet des types de logement. Enfin, dans le cas du modèle 3, seules les maisons en rangée se démarquent des maisons individuelles. Le niveau d'instruction du chef de famille est encore important, mais, en 1978, l'âge du chef est significatif avec un coefficient positif. On observe une certaine amélioration du coefficient  $R^2$  pour 1978, mais celui-ci demeure très faible.

Cette comparaison entre deux années, du point de vue de l'élaboration des politiques, laisse croire que la qualité des maisons individuelles s'est améliorée au Canada. Sur le plan des méthodes, la comparaison montre l'importance d'un choix judicieux de la variable dépendante. Comme on l'a mentionné précédemment, on procédera à de nombreuses autres analyses au moyen des techniques de régression disponibles afin de raffiner ces résultats et de tenir compte du plan de sondage.

Les données sur les dépenses des familles, on l'a vu, ont leurs limites, mais elles contiennent une manne de renseignements importants qu'il vaut la peine d'explorer.

## RÉFÉRENCE

McDonald, Gordon H.G., Ritchie, J.R. Brent et Claxton, John D. (1979), Energy Conservation and Conservation Patterns in Canadian Households: Overview, Behavioral Energy Research Group, 203-2053 Main Mall, University of British Columbia.

examinées avec la plus grande prudence. Pour les besoins de cet article, nous ne noterons que les variables significatives au niveau de 0,01 et au-delà, et nous n'en indiquerons que le signe.

Des variables fictives sont incorporées à la liste des variables indépendantes dans le premier et le second modèle pour la ville, et dans les trois modèles pour le type de logement. Les variables non indiquées sont, en ce qui concerne la ville, Ottawa, et en ce qui concerne le type de logement, la maison individuelle.

En 1974, la taille du logement, certaines villes, les dépenses totales, l'âge du chef et la taille de la famille, ainsi que certains types de logements, sont des variables significatives. Les grandes familles dont les dépenses totales sont élevées et qui vivent à Saint-Jean (I.-N.) dans des maisons individuelles sont celles qui consomment le plus. La consommation est moins forte dans les villes de l'ouest que dans celles de l'est. Elle est par ailleurs plus forte dans les maisons individuelles que dans tous les autres types de logements, bien que pour les duplex, la différence ne soit pas significative lorsque l'on tient compte du nombre de pièces. Les résultats non pondérés sont semblables aux résultats pondérés.

Lorsqu'on prend les dépenses par pièce comme variable dépendante et qu'on retranche le nombre de pièces de la liste des variables indépendantes, le schéma global demeure le même. Toutefois, la taille de la famille n'est plus une variable significative (elle est probablement étroitement liée à la taille du logement seulement), et le niveau d'instruction du chef de famille devient significatif, avec coefficient négatif, c'est-à-dire que moins les personnes sont instruites, plus elles consomment, toutes choses étant égales par ailleurs. Enfin, le duplex devient une variable significative avec coefficient positif, c'est-à-dire qu'une fois soustrait l'effet du nombre de pièces, il se consomme plus d'énergie dans les duplex que dans les maisons individuelles.

Dans le modèle 3, on tient compte des conditions climatiques en incorporant les degrés-jours dans la variable dépendante et en retranchant la liste des villes de l'ensemble des variables indépendantes. Il convient de noter que si la valeur des coefficients baisse de façon si prononcée, c'est parce que le nombre de degrés-jours dans ces villes se situe entre 4000 et 7000. La faible valeur des coefficients ne signifie pas, par conséquent, qu'ils ne sont pas importants. Les dépenses totales demeurent un facteur significatif, tout



éliminant les effets de variables dont le poids est très lourd, mais qui échappent en grande partie ou totalement au contrôle des consommateurs.

Les dépenses consacrées à l'énergie à la maison peuvent être examinées en fonction de certains facteurs, mais comme l'un des principaux facteurs qui conditionnent ces dépenses est la taille de la maison, cet élément peut être ignoré dans la variable d'entrée pour permettre l'examen d'autres facteurs (du point de vue de l'élaboration des politiques). Ainsi, au lieu d'étudier les dépenses totales consacrées à l'énergie, on s'intéresse aux dépenses par pièce. En poussant un peu plus loin, on peut éliminer l'effet des variations climatiques et météorologiques d'une année à l'autre en examinant les dépenses/pièce/degré-jour. Les degrés-jours sont pris en compte par ville et par année. Les données portant sur les degrés jours par année et pour chaque ville furent obtenues d'Environnement Canada. Le tableau 3 montre comment les chiffres varient à mesure que la complexité du facteur étudié s'accroît, encore une fois pour deux années. La comparaison des deux années et les différences dans les taux de changements d'une année à l'autre suggèrent l'importance de raffiner la mesure afin d'améliorer la compréhension du phénomène.

### B) Construction de variables de sortie sommaires pour l'examen de la structure des données - exemple de coefficients de régression.

On présente aux tableaux 4 à 6 certains résultats de l'analyse de régression. Trois modèles sont examinés. D'un modèle à l'autre, la variable dépendante devient plus complexe. Il est ainsi possible de contrôler les facteurs qui ont une incidence marquée sur la consommation d'énergie et de mieux voir dans quelle mesure les autres variables peuvent intervenir de façon significative.

On n'a pas tenté dans ces analyses d'aborder le problème du plan de sondage complexe. Une telle étude sera effectuée plus tard au moyen de la technique de l'interaction de Taylor, et les résultats seront comparés. On présente tout-fois, pour l'année 1974, les résultats obtenus avec un échantillon pondéré et un échantillon non pondéré. Comme on peut le constater, les valeurs des coefficients changent très peu et leur degré de signification demeure le même. À cause des contraintes mentionnées et du fait que des différences très légères peuvent produire des résultats significatifs en raison de la très grande taille des échantillons, ces résultats préliminaires doivent être



très jeunes. Cela tient sans doute au fait que le groupe inclut à la fois des personnes âgées et des ménages monoparentaux (dont les chefs de famille sont probablement des femmes) comprenant de jeunes enfants. On notera aussi que ce groupe possède le moins grand nombre de salariées à plein temps. En 1978, la situation générale demeure la même, sauf que le groupe n'est plus le plus âgé. En fait, le groupe le plus âgé est celui de la case située immédiatement à droite dans la matrice des groupes types. Il semble donc qu'en 1978, les personnes les plus âgées consomment une quantité relativement plus grande d'énergie à la maison. Peut-être ce groupe était-il en meilleure position financière en 1978 qu'en 1974, mais peut-être aussi a-t-il été incapable de contenir les dépenses consacrées à l'énergie en raison de la hausse des prix.

En 1974, les "hippopotames" possèdent aussi les caractéristiques prévues. Il semble qu'il s'agisse de personnes d'âge moyen ayant un grand nombre d'enfants de cinq à seize ans, les familles ayant atteint leur taille maximale. Ce sont eux qui dépendent le plus dans la plupart des catégories. Ils ont en outre le niveau d'instruction le plus élevé. En 1978, toutefois, ce groupe ne se situe plus au sommet sur le plan de l'instruction, ni sur celui des dépenses consacrées au logement. La raison en est peut-être que ceux qui ont les plus grandes maisons et qui sont les plus instruits ont commencé à modifier leurs habitations en vue de réaliser des économies d'énergie.

On note également une évolution chez les "livières". En 1974, ils formaient le groupe le plus jeune, avec de très petites familles. En 1978, ce groupe semble être constitué davantage de jeunes familles ayant de jeunes enfants. Le changement le plus marqué chez ce groupe est la baisse considérable des dépenses consacrées au tabac et aux boissons alcooliques.

Le niveau de signification de ces changements peut être évalué à l'aide de tests statistiques appropriés. Cette analyse avait pour but d'isoler certains groupes types dans la population. On peut ainsi mettre en lumière certaines caractéristiques des modes de vie qui sont très utiles pour adapter les programmes d'économie d'énergie à chaque groupe.

D'autres analyses pourraient porter non pas sur les niveaux des dépenses, mais sur les pourcentages correspondants. Cela permettrait de dresser le profil de ceux qui les dépenses consacrées à l'énergie représentent le fardeau le plus lourd.

(11) On peut produire des variables d'entrée complexes continues qui

Nous avons choisi les suivantes :  
 A) Construction de variables d'entrée complexes, afin de s'en tenir à l'étude des facteurs les plus significatifs.  
 (1) On a produit des variables complexes discontinues en combinant les catégories d'énergie à la maison et pour le transport, mais en ne considérant que ces deux catégories de dépenses continues. On a formé des tranches de ces dernières afin de former des groupes types de ménages, dont on peut ensuite étudier les caractéristiques pour mettre en lumière leurs différences. Pour créer ces groupes, on a divisé les dépenses des deux catégories en quartiles, on a combiné les deux quartiles centraux, puis on a formé une matrice à neuf cases à partir des deux ensembles de trois tranches résultantes (voir le tableau 1; source : McDougall, Ritchie et al., 1974). Il est intéressant de comparer entre elles et avec la case centrale les cases qui se trouvent aux angles de cette matrice. Ces groupes types ont été utilisés dans le cadre d'une étude antérieure réalisée pour le ministère de la Consommation et des Corporations; la comparaison entre les résultats obtenus avec les données sur les dépenses des familles et les autres ensembles de données utilisés dans l'étude du MCC se révèle donc spécialement intéressante. On voudra aussi étudier comment les caractéristiques de ces groupes évoluent avec le temps. Par exemple, est-ce que la catégorie des "souris d'entrée" demeure formée de Canadiens à faible revenu (faible revenu) ou décèle-t-on une tendance vers l'adoption de modes de vie moins économes. Au tableau 2, on compare les caractéristiques de trois groupes de "souris d'entrée", littéraires et hipopodames) pour deux années différenciées. On donne d'abord, dans le cas des "souris d'entrée", des renseignements relatifs à une gamme possible de variables d'analyse, pour les années 1974 et 1975. Afin de simplifier la présentation, seul le rang de la case parmi l'ensemble des groupes types est indiqué. Traditionnellement, ceux qui consomment la moins d'énergie sont ceux qui, en général, sont les plus dépourvus: revenus les plus bas, faible niveau d'instruction, âge avancé. On peut voir que les "souris d'entrée" possèdent bel et bien ces caractéristiques en 1974. Ce groupe type se situe également au niveau le plus faible pour toutes les catégories de dépenses indiquées. Bien qu'on y retrouve les personnes les plus âgées, le groupe ne vient pas au rang le plus bas pour le nombre d'enfants

locataires d'appartements situés dans des immeubles à compteur central et les chaudières.

Certains chercheurs ont imputé des valeurs à ces ménages en se fondant sur leurs loyers, mais nous avons choisi de ne pas procéder ainsi et de restreindre plutôt notre étude aux ménages qui sont en mesure de surveiller et de contrôler leur propre consommation d'énergie. Il s'agit en effet des groupes de consommateurs auxquels s'adressera tout programme gouvernemental visant à modifier la consommation.

Plusieurs facteurs font qu'il est difficile d'étudier et d'infléchir les schémas de consommation d'énergie des ménages :

- les caractéristiques des biens : elles peuvent limiter la capacité du ménage d'acquies rapidement et augmenter le coût des changements (ces caractéristiques comprennent la taille de la maison, le nombre et le type d'appareils ménagers, la taille et le nombre de véhicules, etc.) On note dans certaines études que les caractéristiques des maisons peuvent expliquer à elles seules 24% de la consommation d'énergie au foyer. La taille du ménage peut être incluse au nombre des caractéristiques des biens.
- la possibilité de changement : il n'est pas toujours possible de changer la quantité et les types de combustibles utilisés, car on n'a pas toujours accès, selon la situation ou les montants à investir, à la technologie ou aux combustibles voulus : le chauffage au gaz naturel, par exemple, est inaccessible aux résidents des régions rurales : on ne peut pas, dans ce cas, modifier instantanément le type de combustible utilisé.
- facteurs exogènes : ils influent sur la quantité d'énergie nécessaire pour obtenir un même rendement dans des situations différentes (au nombre de ces facteurs figurent les conditions météorologiques, les distances à parcourir en ville, etc.).

### 3. COMBINAISON DES DONNÉES D'ENTRÉE ET OPTIMISATION DES RÉSULTATS

Devant un ensemble de données si complexe et un sujet si difficile et si polyvalent, il faut tâcher de réduire les flux d'information, tout en augmentant le contenu informatif de chaque facteur qui intervient dans les analyses,

documents. L'exacitude des données ainsi obtenues pose des problèmes au niveau des individus, mais ces problèmes sont atténués lorsqu'on dispose d'échantillons très grands. Dans la plupart des enquêtes indépendantes, toutefois, il en coûterait trop cher de former des échantillons aussi vastes. Heureusement, les données de l'enquête sur les dépenses des familles provien-

nent d'échantillons de très grande taille.

Il n'est arrivé qu'une seule fois au Canada qu'on tente dans une enquête indépendante d'effectuer une vérification indépendante de dossiers, en se procurant les documents relatifs à chaque ménage auprès des fournisseurs, avec la permission du chef de ménage. Toutefois, on n'a pu obtenir des dossiers sur la consommation d'électricité que pour moins de la moitié de l'échantillon, et sur la consommation de mazout et de gaz naturel, pour à peu près le tiers de l'échantillon seulement. Ce procédé de vérification des dossiers assure un type élevé d'exacitude, élimine les problèmes liés au besoin de se rappeler des événements passés, surtout s'ils remontent à une période éloignée, et évite l'introduction d'un biais de réponse. Il est toutefois impossible, en pratique, de le mettre en oeuvre pour de vastes échantillons couvrant l'ensem-

ble du pays. Rien que l'enquête sur les dépenses des familles fasse appel à la mémoire des répondants, les renseignements sur les dépenses consacrées à l'énergie ne sont probablement pas entachés d'un biais aussi grand que dans une enquête portant expressément sur la consommation d'énergie, car les répondants ne sont pas sensibilisés à l'objet de l'enquête. Par ailleurs, les données ont été recueillies selon les mêmes méthodes avant et après la crise de l'énergie, ce qui réduit encore la possibilité de biais de réponse. Les données de cette enquête offrent donc une occasion unique d'observer un ensemble très vaste d'échantillons pendant une période cruciale.

Les données posent quand même certains problèmes, dont certains sont attribuables à la méthode d'échantillonnage et d'autres provenant de la complexité inhérente à toute enquête sur la consommation d'énergie. En raison des renseignements apportés aux catégories de dépenses et à leur contenu, en particulier aux catégories autres que l'énergie, il nous a fallu manipuler considérablement les données afin d'en assurer l'uniformité d'une période à l'autre. Il est impossible de connaître les dépenses consacrées à l'énergie à la maison par les familles qui ne paient pas cette énergie directement, par exemple les



Corporations et de l'Énergie, des Mines et des Ressources, où l'on continue de mener des programmes de recherche actifs sur l'utilisation et la conservation de l'énergie par les consommateurs. La structure du projet tient compte des intérêts, des orientations et des contraintes de ces ministères.

Par ailleurs, au cours des cinq dernières années, un groupe international de spécialistes des sciences sociales a entrepris un programme de recherche et d'échange d'information sur les comportements des consommateurs et les utilisations de l'énergie. À titre de membre du groupe, l'auteur est bien au fait des problèmes et des perspectives examinés par ce groupe, ainsi que de l'état actuel de ses travaux et de ses techniques de recherche.

## 2. PROBLÈMES LIÉS À LA RECHERCHE SUR LES QUESTIONS ÉNERGÉTIQUES

Le problème principal qu'on rencontre dans les études sur la consommation d'énergie, c'est peut-être celui d'obtenir des mesures raisonnablement fiables de la consommation à partir d'échantillons suffisamment vastes et représentatifs. Un chercheur serait comblé si, en outre, il pouvait disposer de telles données pour une certaine période, en particulier si celle-ci englobait le fameux choc pétrolier de 1973. Les données sur les dépenses des familles recueillies par la Division des revenus et des dépenses des familles de Statistique Canada correspondent assez bien à cette description pour au moins soulever l'intérêt du chercheur. Il s'agit de données provenant d'une série d'enquêtes rétrospectives réalisées pour les années 1969, 1972, 1974, 1976, 1978 et 1982. La période intéressante est donc couverte, l'échantillon est vaste et les méthodes d'échantillonnage employées garantissent une bonne représentativité pour les secteurs étudiés, habituellement les centres urbains. En outre, les données comprennent un grand nombre d'autres variables utiles dans toute étude de la consommation d'énergie, par exemple le mode d'occupation des logements, certaines caractéristiques des logements, le nombre de propriétaires de véhicules et d'appareils ménagers, les caractéristiques des familles, les dépenses consacrées à d'autres catégories de biens et de services de consommation, etc.

Dans la plupart des enquêtes sur les dépenses de consommation, on demande aux répondants de fournir des renseignements de mémoire ou par référence à des

ANALYSE DES DÉPENSES CONSACRÉES À L'ÉNERGIE

Louise A. Heslop<sup>1</sup>

À l'aide des données chronologiques de l'enquête sur les dépenses des familles, on effectue actuellement des analyses des dépenses de consommation d'énergie pour la maison et le transport, pour les années 1969 à 1982. Dans le présent article, on décrit brièvement certaines des méthodes d'analyse utilisées, on présente des résultats sommaires et on s'intéresse aux changements survenus dans la consommation dans la mesure où ils peuvent influencer sur l'orientation des politiques. Devant un ensemble de données si complexe et un sujet si difficile et si polyvalent, il faut tâcher de réduire les flux d'information, tout en augmentant le contenu informatif de chaque facteur qui intervient dans les analyses, tant en entrée qu'en sortie.

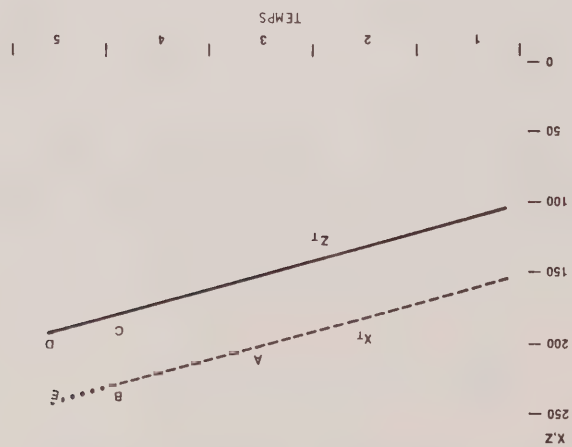
1. LA QUESTION ÉNERGÉTIQUE

Pour certains, le problème de la conservation de l'énergie ne se pose plus. Il n'y a pas de pénurie (peut-être n'y en a-t-il jamais eue), et les prix se sont stabilisés.

La question énergétique a dominé l'actualité des années 1970, bouleversant profondément l'ordre économique mondial et engendrant des affrontements à l'échelle internationale. Au pays, cette question a eu de lourds effets sur les relations fédérales-provinciales, sur les liens entre le monde des affaires et le gouvernement et sur les budgets familiaux. Elle a entraîné une restructuring de l'industrie manufacturière, du secteur de l'automobile, etc.

Bien qu'apparemment relégués au second plan, les problèmes liés à la consommation et aux prix de l'énergie demeurent prioritaires pour les consommateurs, les entreprises et le gouvernement. La conservation de l'énergie ne soulève plus les passions, mais elle demeure un problème crucial. La recherche dont fait état cet article a été organisée en consultation avec les responsables des politiques aux ministères de la consommation et des





**Figure 5:** Continuité entre la série ajustée aux jalons (ligne en tirets) et la série préliminairement ajustée (pointillée) et discontinuité BC entre les séries ajustées (tirets) et non ajustée (continue).

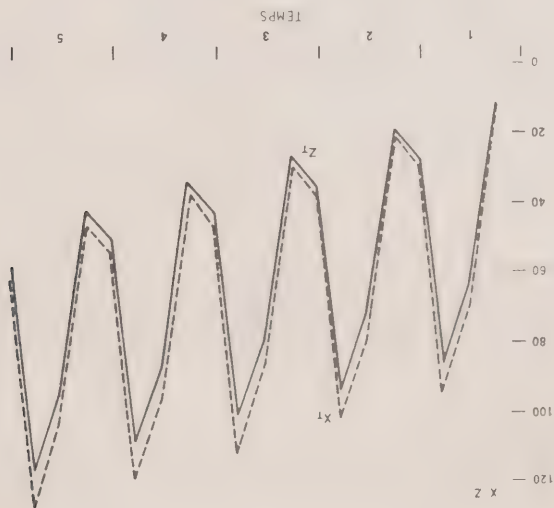


Figure 3: Série originale (ligne continue) et série ajustée aux jalons (tirets) selon la variante proportionnelle de la méthode d'étalonnage proposée dans ce travail (en régime d'écartes annuels proportionnels constants).

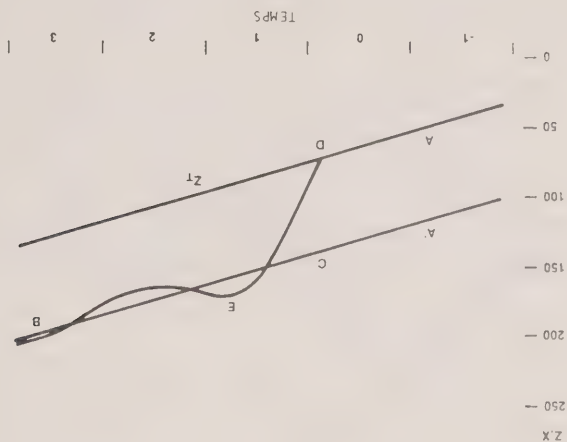
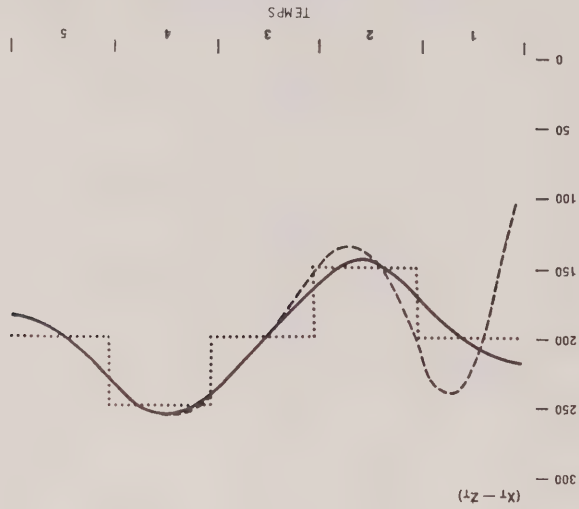
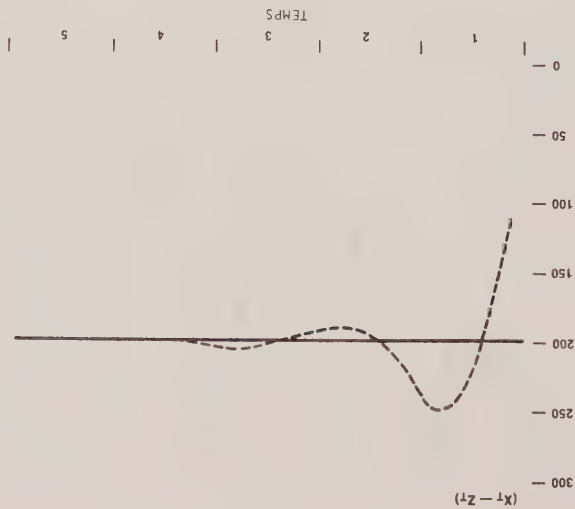


Figure 4: Séries ajustées aux jalons selon la méthode de Denton, en l'absence de jalons pour les années -1 et 0 (courbe ADER) et en présence de jalons annuels et d'ajustement préalable pour les années -1 et 0 (A'CDER); et selon la méthode proposée dans ce travail, appliquée à la manière d'une moyenne mobile, en présence de jalons annuels pour les années -1 et 0 (A'CB).

**Figure 2:** Corrections ( $x_t - z_t$ ) apportées à la série non ajustée aux jalons selon la méthode de Denton (Ligne en tirets) et selon la méthode d'étalement proposée dans ce travail (continue), en régime d'écartes annuels moyens variables (pointillée).



**Figure 1:** Corrections ( $x_t - z_t$ ) apportées à la série non ajustée aux jalons selon la méthode de Denton (Ligne en tirets) et selon la méthode d'étalement proposée dans ce travail (continue), en régime idéal d'écartes annuels constants.



- [111] Denton, F.T. (1971), "Adjustment on Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization," J.A.S.A., Vol. 66, No. 333, pp. 99-102.
- [112] Fernandez, R.B. (1981), "A Methodological Note on the Estimation of Time Series," Review of Economic and Statistics, Vol. 63, pp. 471-476.
- [113] Friedman, M. (1962), "The Interpolation of Time Series by Related Series," J.A.S.A., Vol. 57, No. 300, pp. 729-757.
- [114] Gleijser, H. (1966), "Une méthode d'évaluation de données mensuelles à partir d'indices trimestriels ou annuels," Cahiers Economiques de Bruxelles, No. 19, 1er trimestre, pp. 45-64.
- [115] Helfand, S.D. Monsour, N.J. Traeger, M.L. (1978), "Historical Revision of Current Business Survey Estimates," U.S. Bureau of the Census, (Document de recherche).
- [116] Huot, G. (1975), "Quadratic Minimization of Monthly Estimates to Annual Totals," Séries chronologique recherche et analyse, Statistique Canada, Document de recherche 75-11 010E.
- [117] Lisman, J.H.C., Sandee, J. (1964), "Derivation of Quarterly Figures from Annual Data," Applied Statistics, Vol. 13, No. 2, pp. 87-90.
- [118] Smith, P. (1977), "Alternative Method for Step Adjustment," Groupes des analyses de conjoncture, Statistique Canada, document de recherche.
- [119] Gormeyer, W.H., Jansen, R., Lauter, A.S. (1976), "Estimating Quarterly Values from Annually Known Variables in Quarterly Relationships," J.A.S.A., Vol. 71, No 355, pp. 588-595.
- [120] Wilcox, J.A. (1983), "Disaggregating Data Using Related Series," Journal of Business and Economic Statistics, Vol. 1, No 3, pp. 187-191.

- [2] Bassie, B.L. (1939), "Interpolation Formulae for the Adjustment of Index Numbers," Proceedings of the Annual Meetings of the American Statistical Association.

- [3] Root, J.C.G., Feibes, W., Lisman, J.H.C. (1967), "Further Methods of Derivation of Quarterly Figures from Annual Data," Applied Statistics, Vol. 16, No. 1, pp. 65-75.

- [4] Cholette, P.A. (1978), "Comparaison et évaluation de quelques méthodes d'ajustement de séries infra-annuelles aux repères annuels," Séries chronologique recherche et analyse, Statistique Canada, document de recherche 78-03-001B.

- [5] Cholette, P.A. (1979a), "Adjustment Methods of Sub-Annual Series to Yearly Benchmarks," Proceedings of the Computer Science and Statistics, 12th Annual Symposium on the Interface, J.F. Gentleman Ed., University of Waterloo, pp. 358-36.

- [6] Cholette, P.A. (1979b), "A Note on 'Freezing' Past Estimates when Benchmarking," Séries chronologique recherche et analyse, Statistique Canada, document de recherche 79-06-002E.

- [7] Cholette, P.A. (1982), "Programme d'ajustement quadratique minimum (PAQM-I) de séries aux totaux annuels - Manuel des utilisateurs," Séries chronologique recherche et analyse, Statistique Canada, 82-11-003B.

- [8] Cholette, P.A. (1983), "Étalonnage de séries en régime de jalons bi-annuels et de connaissance du point d'arrivée," Séries recherche et analyse, Statistique Canada, document de recherche 83-05-002B.

- [9] Chow, G.C., Lin, An-Ioh (1971), "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series," Review of Economics and Statistics, Vol. 53, No. 4, pp. 372-375.

- [10] Daum, E.B. (1977), "Comparison of Various Interpolation Procedures for Benchmarking Economic Time Series," Séries chronologique recherche et analyse, Statistique Canada, document de recherche 77-05-006E.

mouvements des séries apparentées (et la série trouvée satisfait aux contraintes annuelles). Fernandez (1981) note que la méthode de Chow et Lin peut produire des discontinuités de mouvements entre les années. Il propose alors une synthèse des méthodes de Chow et Lin et de Denton (1971). La méthode combinée élimine les discontinuités inter-annuelles mais repose également sur l'hypothèse  $x_0 = z_0$ . Comme illustré plus haut, cette hypothèse engendre souvent des fluctuations artificielles dans la série calculée. Nous aurions tendance à croire qu'il est possible de s'abstenir de cette hypothèse dans le cas de la méthode de Fernandez comme pour la méthode de Denton.

## 7. RÉSUMÉ ET CONCLUSIONS

Denton (1971) voulait garder les séries originales et ajustées aux jalons annuels aussi parallèles que le permettaient les écarts annuels. Ce travail a impliqué une modification de la méthode d'étalement qui rend les séries originales et ajustées plus parallèles l'une à l'autre que ce n'est le cas avec la méthode originale. Cette amélioration vaut autant pour les variantes additives que proportionnelles. Nous pensons que la méthode généralisée multipliée de Fernandez pourrait être améliorée dans le même sens.

On peut tout aussi facilement adapter la méthode proposée aux séries de flux, de stock et d'indice.

Avant de procéder à des comparaisons inter-temporelles entre les données ajustées et les nouvelles données courantes, il est essentiel d'ajuster aux jalons de manière préliminaire les données courantes (de la façon proposée). La mise en oeuvre suggérée, à la façon d'une moyenne mobile quinquennale, préservera automatiquement intactes les estimations passées deux années de révision.

## RÉFÉRENCES BIBLIOGRAPHIQUES

- [1] Baldwin, A. (1978), "New Benchmarking Algorithms using Quadratic Minimization," National Product Division, Statistique Canada, Document de recherche.



données non ajustées (accompagnées de leurs jalons annuels). Pour des méthodes (avec leurs jalons). La soumission de données ajustées provoquera généralement un mouvement saisonnier artificiel dans la série ajustée résultante (Cholette, 1978, section 6b).

### 6.3 Étalonnage préliminaire des données courantes

Finalement une dernière remarque s'impose. Pendant une année (incomplète) en cours, on ne peut pas calculer des taux de croissance, par exemple, entre le segment de la série ajustée aux jalons (AB) avec le segment non ajusté (CD). Ce faire produit généralement une discontinuité BC entre les deux segments AB et CD, comme illustré dans la figure 5 par la courbe ABCD.

Deux issues s'offrent alors. Premièrement, on fait les comparaisons inter-temporelles en se basant seulement sur les données non ajustées. Deuxièmement, on effectue un ajustement préliminaire des données courantes, en répartition de l'année incomplète courante dans la fonction objective (4) (ou 12) produirait des valeurs ajustées préliminaires identiques.) On peut ensuite comparer le segment ajusté AB avec le segment préliminairement ajusté AC, comme illustré dans la figure 5 par la courbe ABC. Nous favorisons cette deuxième alternative.

### 6.4 Rapport avec les autres méthodes

On pourrait qualifier d'univariées les méthodes d'étalonnage de Denton (1971), de Denton modifiée (présentée ici), de Glejser (1966), de Boot, Feibes et Lisman (1967), de Lisman et Sandee (1964), et de Bassie (1939). En effet, ces méthodes ne font intervenir que la série considérée et ses jalons annuels dans le processus d'étalonnage. Les méthodes de Friedman (1962), Chow et Lin (1971) Somermeyer, Jansen et Louter (1976), et de Wilcox (1983) sont au contraire multivariées. Des séries auxiliaires y sont utilisées dans le calcul de la série recherchée.

Par exemple, Chow et Lin (1971) ont proposé une méthode pour obtenir la série infra-annuelle désirée à partir de totaux annuels et de série apparentes. Le mouvement de la série résultante est le plus semblable possible aux

Si on laisse ces années intactes parce qu'elles n'ont jamais eu de repères annuels, la solution proposée par Denton est défendable: il n'en découle aucune correction pour les années -1 et 0: et, de petites corrections graduellement introduites au début de l'année 1. (On se souvient que  $x_0 = z_0$  implique la minimisation de la première correction.) La série ajustée résultante est donc continue, comme illustré à la figure 4 par la courbe ADEB.

Par contre, si on laisse intactes les premières années parce qu'elles ont été ajustées aux jalons et qu'on les considère maintenant comme "historiques", nous ne sommes pas d'accord avec l'hypothèse  $x_0 = z_0$ . Généralement en effet, celle-ci produira une discontinuité entre les années 0 et 1 comme illustré dans la figure 4 par la courbe A'CEB. Les années -1 et 0 ont déjà reçu des corrections de taille voisine de CD, tandis que le début de l'année 1 reçoit des corrections les plus petites possibles.

Pour rendre immuables les données historiques après un certain nombre d'années, deux solutions sont possibles. La première consiste à spécifier explicitement la contrainte d'immuabilité dans la fonction objective qui devient

$$p(x) = ((x_1 - z_1) - (x_0 - z_0))^2 + \sum_{t=2}^T (\Delta(x_t - z_t))^2, \quad (16)$$

où  $(x_0 - z_0)$  est connu et égal à la dernière correction utilisée pour l'année historique 0. Cette correction est généralement différente de zéro (Cholette, 1979, 1983). Cette spécification équivaut à déterminer le point de départ de la courbe de correction.

Une deuxième solution, moins spécifique mais aussi efficace, consiste à appliquer la méthodologie déjà proposée dans ce travail (version additive ou proportionnelle) à la manière d'une moyenne mobile se mouvant annuellement. Dans un intervalle quinquennal d'application, les estimés deviennent automatiquement définitifs après deux années de révision: et, après une année, en régime triennal (Cholette, 1978, section 6 a; 1974, 4.3). La série ajustée aux données résultante est également continue, comme illustre la courbe A'CB de la figure 4.

## 6.2 Mise en oeuvre

Les praticiens de l'étalonnage ont tendance à soumettre au programme d'étalonnage les années de données déjà ajustées aux jalons suivie d'une année de

$$\begin{pmatrix} \bar{g} \\ \bar{x} \end{pmatrix} = \begin{pmatrix} \bar{Z}^{-1} \bar{A} & \bar{Z}^{-1} \bar{B} \\ \bar{0} & \bar{I} \end{pmatrix}^{-1} \begin{pmatrix} \bar{Z}^{-1} \bar{A} & \bar{Z}^{-1} \bar{B} \\ \bar{0} & \bar{I} \end{pmatrix} \begin{pmatrix} \bar{I} \\ \bar{Z} \end{pmatrix} = \begin{pmatrix} \bar{I} \\ \bar{W} \bar{x} \end{pmatrix} \begin{pmatrix} \bar{I} \\ \bar{W} \bar{y} \end{pmatrix} \quad (15)$$

Contrairement à ceux de la variante additive, on doit cependant calculer les poids  $W^x$  de la solution proportionnelle pour chaque série et même chaque intervalle d'application d'une série donnée.

## 5. SÉRIES DE STOCKS ET D'INDICE

Les variantes additive et proportionnelle de la méthode présentée ci-dessus sont conçues pour des séries de flux, dont la valeur annuelle correspond à la somme des valeurs infra-annuelles. On peut très facilement adapter les solutions trouvées aux séries de stock ou cumulatives, dont la valeur annuelle n'est associée qu'à une seule valeur infra-annuelle (habituellement celle du dernier mois): ainsi qu'aux séries d'indice, dont la valeur annuelle correspond à la moyenne des valeurs infra-annuelles. Pour une série trimestrielle cumulative par exemple, il suffit tout simplement de redéfinir le vecteur  $\bar{J}$ , composante de  $\bar{B}$ , comme ceci

$$\bar{J}^{1 \times 4} = [0 \ 0 \ 0 \ 1]$$

et, pour une série mensuelle d'indice comme ceci

$$\bar{J}^{1 \times 12} = [1/12 \ 1/12 \ 1/12 \ \dots \ 1/12]$$

## 6. DISCUSSION

### 6.1 Données historiques

Il y a selon nous beaucoup de confusion quant à l'interprétation de l'hypothèse  $x_0 = z_0$  de Denton. L'auteur écrit à ce sujet: "On suppose qu'il n'y a pas d'ajustement à faire à la série originale pour les années extérieures à l'intervalle allant des années 1 à m inclusivement." (p. 100, au dessus de l'équation 3.2, notre traduction).

section 2. la fonction objective minimise toujours la somme des différences  
 d'attributions de pente entre les séries infra-annuelles originale et recherchée  
 ( $z_t$  et  $x_t$ ). Mais chaque terme de la somme est pondéré par la valeur de l'ob-  
 servation infra-annuelle correspondante:

$$p(x) = \sum_{t=2}^T (\Delta(x_t - z_t)/z_t)^2 = \sum_{t=2}^T (\Delta(x_t/z_t))^2. \quad (12)$$

Cette variante convient aux séries à forte saisonnalité, lorsqu'on juge que  
 les mois de creux saisonnier ne peuvent être aussi responsables de l'écart  
 annuel que les mois de sommet saisonnier. La taille de chaque correction est  
 proportionnelle au niveau de l'observation, comme illustré dans la figure 3.  
 Les observations faibles échappent de corrections plus petites que les observa-  
 tions saisonnièrement fortes, même si les corrections proportionnelles  $z_t/z_t$   
 minimisées sont aussi constantes que les écarts annuels le permettent. A  
 l'inverse, qu'avec la variante proportionnelle, toutes les observations doivent  
 être positives et que les valeurs ajustées seront aussi toutes positives.

On peut également démontrer (Cholette, 1978, section 3; 1979, 3) que la  
 variante proportionnelle est une approximation linéaire de la méthode forte-  
 ment non linéaire de préservation des taux de croissance (Smith, 1977; Helfand  
 et al., 1978) qui aurait la fonction objective suivante:

$$p(x) = \sum_{t=2}^T (x_t/x_{t-1} - z_t/z_{t-1})^2 \quad (13)$$

L'approximation est exacte en régime d'écarts annuels proportionnels  
 car nous sur l'intervalle d'estimation.  
 En algèbre linéaire, la fonction objective contrainte associée à la méthode  
 proportionnelle s'écrit

$$u(\bar{x}, \bar{g}) = (\bar{x} - \bar{z})' \bar{Z}^{-1} \bar{A} \bar{Z}^{-1} (\bar{x} - \bar{z}) - 2 \bar{g}' (\bar{y} - \bar{B}' \bar{x}), \quad (14)$$

où  $\bar{Z}^{-1}$  est une matrice diagonale dont les éléments sont  $1/z_1, 1/z_2, \dots$ . La  
 solution à la même structure que la variante additive ( $\bar{Z}^{-1} \bar{A} \bar{Z}^{-1}$  remplaçant  $\bar{A}$   
 dans (11)) et s'écrit:

$$\begin{pmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} \bar{A} & \bar{B} \\ \bar{B}' & \bar{0} \\ \bar{A} & \bar{0} \end{pmatrix}^{-1} \begin{pmatrix} \bar{0} \\ \bar{0} \\ \bar{I} \end{pmatrix} \begin{pmatrix} \bar{y} \\ \bar{z} \end{pmatrix} = \frac{1}{W} \begin{pmatrix} (n+m) \times (n+m) \\ \bar{W} \end{pmatrix} \begin{pmatrix} \bar{y} \\ \bar{z} \end{pmatrix} \quad (10)$$

La substitution de l'identité  $\bar{y} = \bar{B}'\bar{z} + \bar{r}$ ,  $\bar{r}$  renfermant les écarts annuels, donne

$$\begin{pmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} \bar{A} & \bar{B} \\ \bar{B}' & \bar{0} \\ \bar{A} & \bar{0} \end{pmatrix}^{-1} \begin{pmatrix} \bar{0} \\ \bar{0} \\ \bar{I} \end{pmatrix} \begin{pmatrix} \bar{y} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} \bar{I} & \bar{0} \\ \bar{0} & \bar{W} \end{pmatrix} \begin{pmatrix} \bar{y} \\ \bar{z} \end{pmatrix} \quad (11)$$

Cette reformulation de la solution réduit le temps de calcul dans l'application des poids calculés comparativement à la formulation (10). À remarquer aussi qu'une fois obtenus, les poids  $W^x$  peuvent servir pour un nombre quelconque de séries ayant le même nombre d'observations. Nous recommandons en outre (Chollette, 1978, section 6: 1979, 4.3) de calculer  $W^x$  pour un intervalle quinquennal et de l'utiliser à la manière d'une moyenne mobile (se mouvant d'une année à la fois) pour les séries de cinq ans et plus. En plus d'économiser les calculs, ce procédé n'engendre que deux révisions des estimés (ceteris paribus) lorsque d'autres années d'observations s'ajoutent à la série.

Denton résout l'inversion de l'équation (10) par parties. Cela est impossible ici, car la matrice  $A$  est singulière. Par contre, l'ensemble de la matrice n'est pas singulier et s'inverse.

La méthode exposée ici utilise en fait la solution que Boot, Feibes et Lisman (1967) proposaient pour interpoler entre des données annuelles en l'absence d'information infra-annuelle. La solution (11) consiste exactement à interpoler entre les écarts annuels avec la méthode de ces auteurs et à ajouter les estimés obtenus (les corrections) à la série infra-annuelle originale.

#### 4. VARIANTE PROPORTIONNELLE

La méthode proportionnelle maintenant présentée est aussi une variante de la méthode proportionnelle de Denton, dont on a retenu  $x_0 = z_0$ . Comme dans la

En algèbre linéaire, la fonction objective contrainte s'écrit  
 soujette aux mêmes contraintes de l'équation (2).

$$(5) \quad u(\bar{x}, \bar{q}) = (\bar{x} - \bar{z})' \bar{A} (\bar{x} - \bar{z}) - 2 \bar{q}' (\bar{y} - \bar{B}' \bar{x}),$$

où les vecteurs matrices impliqués valent.

$$(6) \quad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} + \frac{1}{z} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{bmatrix} = \frac{1}{q} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$(7) \quad \frac{A}{n \times n} = \frac{D}{(n-1) \times n}, \quad \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{bmatrix}$$

$$(8) \quad \frac{B}{n \times m} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad \frac{1}{j} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad (n = km).$$

Le vecteur  $\bar{q}$  contient les multiplicateurs de Lagrange. Les variables  $n$  et  $k$  désignent respectivement le nombre d'observations et le nombre d'années dans la série et le nombre de mois par année

Les équations normales associées à la fonction objective (5) sont

$$(6) \quad \begin{aligned} du/d\bar{x} &= (\bar{A} + \bar{A}') (\bar{x} - \bar{z}) + 2 \bar{B}' \bar{q} = 0 \\ du/d\bar{q} &= 2 (\bar{B}' \bar{x} - \bar{y}) = 0 \end{aligned}$$

et débouchent sur la solution



$$(1) \quad p(x) = \sum_{t=1}^T (\Delta x_t - \Delta z_t)^2 = \sum_{t=1}^T (\Delta(x_t - z_t))^2, \quad x_0 = z_0,$$

où  $z_t$  représente la série infra-annuelle originelle au temps  $t$ . Cette fonction est minimisée soumise aux contraintes d'égalité entre les sommes annuelles des valeurs obtenues et les jalons disponibles  $y_t$ :

$$(2) \quad \sum_{i=1}^k x_t = y_t, \quad t = 1, 2, \dots, m.$$

où  $k$  est le nombre de "mois" par année.

Denton justifie l'hypothèse  $x_0 = z_0$  en prétendant qu'il est légitime de supposer l'égalité des dernières valeurs observées et ajustées, antérieures à l'intervalle d'estimation. La fonction objective (1) signifierait donc que la série ajustée  $x_t$  devrait avoir la même pente que la série originelle  $z_t$ : et, que par conséquent la pente des différences entre les deux séries devrait être minimisée (soumise aux contraintes). Mais, on peut écrire la fonction objective (1) de la manière suivante en substituant  $x_0 = z_0$ :

$$(3) \quad p(x) = (x_1 - z_1)^2 + \sum_{t=2}^T (\Delta(x_t - z_t))^2.$$

Cette transformation met clairement en évidence que l'hypothèse  $x_0 = z_0$  implique la minimisation de la taille de la première correction. Comme il n'est pas aux figures 1 et 2, la minimisation de la première correction tire la courbe de correction vers zéro en début de série. Cela produit une ondulation dans la première année qui se répercute sur les autres années. Cette ondulation dans les corrections empêche, par définition, le parallélisme maximum des séries observées et ajustées.

La spécification proposée ici s'abstient tout simplement de l'hypothèse  $x_0 = z_0$  et donne la fonction objective suivante:

$$(4) \quad p(x) = \sum_{t=2}^T (\Delta(x_t - z_t))^2.$$

Voici, suivons le modèle de Ehrenberg (1982) pour la présentation des textes statistiques. Le lecteur se verra exposé les illustrations et résultats d'abord; et, les détails méthodologiques, ensuite.

## 2. ILLUSTRATION DE LA MÉTHODE

La figure 1 montre les corrections  $(x_t - z_t)$  apportées à la série originale selon la solution additive (avec premières différences) de Denton et selon la solution correspondante proposée dans ce travail. Puisque ces corrections doivent être ajoutées à la série infra-annuelle originale  $z_t$ , la série ajustée doit tout à fait parallèle à la série originale si et seulement si les corrections sont constantes. Dans la figure, cela se produit seulement pour les corrections associées à la méthode proposée dans ce travail.

La figure 1 montrait un cas trivial et idéal qui admettait la solution des différences constantes: tous les écarts annuels moyens, différences entre les jalons annuels et les totaux annuels de la série originale (divisées par le nombre de mois par année), étaient constants. La figure 2 propose un cas plus réaliste, où les cinq écarts annuels moyens fluctuent autour de 200. Comme dans le premier exemple, les corrections calculées selon la méthode présentée ici sont beaucoup plus constantes, surtout dans la première année.

Comme expliqué plus bas, la méthode de Denton minimise non seulement le débiais dans les corrections (pour les rendre les plus constantes possible) mais aussi la taille de la première correction. Cela se vérifie dans les figures 1 et 2, où les premières corrections avoisinent zéro. Par contre, la solution alternative minimise seulement le changement dans les corrections. Cependant, cela consiste à tracer, à travers les écarts annuels moyens, une courbe qui soit la plus plate possible et qui recouvre aussi les mêmes surfaces annuelles que les écarts annuels moyens.

## 3. CONSERVATION DU PARALLÉLISME DES SÉRIES ORIGINALE ET AJUSTÉE

Reprenant la formulation additive avec premières différences de Denton pour sa notation, la série recherchée  $x_t$  minimise la fonction objective suivante

## L'AJUSTEMENT DES SÉRIES INFRA-ANNUELLES AUX RÉPÈRES ANNUELS

Pierre A. Cholette<sup>1</sup>

Ce travail propose une modification de la méthode d'étalement de Denton (1971) pour l'ajustement des séries infra-annuelles aux taux annuels. Ces totaux proviennent de source plus fiable et consistent en des repères ou jalons annuels. La série ajustée selon la méthode modifiée s'avère plus parallèle à la série non ajustée que ce n'est le cas avec la méthode originale. Des variantes additives et proportionnelles de la méthode sont exposées. Elles s'adaptent facilement aux séries de flux, de stock et d'indice. On trouve aussi quelques recommandations relatives à l'étalement préliminaire des données courantes et à la gestion des estimés "historiques" de la série.

## 1. INTRODUCTION

Dans un grand nombre de cas, le statisticien obtient des données infra-annuelles d'une série à partir d'une source de données (telle un échantillon); et, les valeurs annuelles repères correspondantes à partir d'une autre source de données plus fiable (telle un recensement). Les totaux annuels des observations infra-annuelles ne sont généralement pas égaux aux valeurs annuelles repères. De telles séries nécessitent l'ajustement aux jalons annuels, c'est-à-dire l'étalement.

La solution proposée par Denton (1971) (et généralisée par Fernandez en 1981) consiste à trouver une série infra-annuelle qui épouserait le plus possible le mouvement de la série infra-annuelle disponible et dont les sommes (ou moyennes) annuelles correspondraient aux repères annuels plus fiables. Le niveau de la série résultante serait ainsi donné par les repères annuels, tandis que son mouvement serait gouverné par la série infra-annuelle originale. En d'autres mots, la série ajustée devrait être la plus parallèle possible à la série originale, tout en satisfaisant aux jalons annuels. Ce travail recommande une modification à la spécification de Denton qui rend les séries originales et ajustées encore plus parallèles.

Figure 5: Taux de sélection observés, Enquête canadienne sur la santé et l'invalidité, janvier 1983, hommes 15-64.

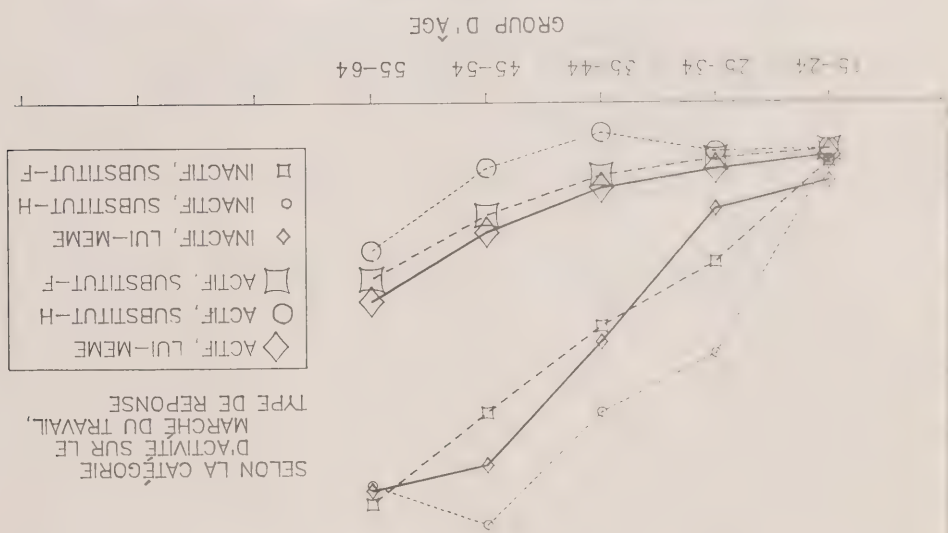


Figure 6: Taux de sélection prédits, Enquête canadienne sur la santé et l'invalidité, janvier 1983, hommes 15-64.

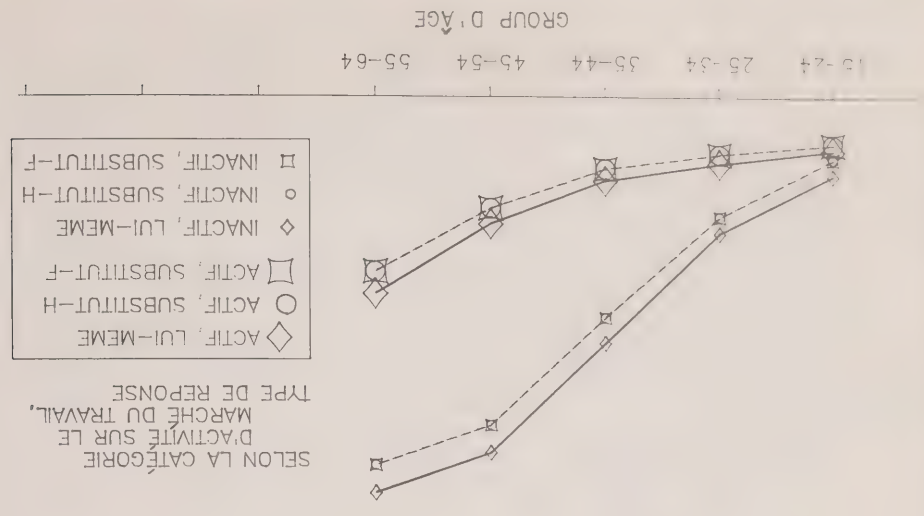


Figure 3: Moyenne observée du nombre annuel de visites chez les médecins par habitant, États-Unis 1973.

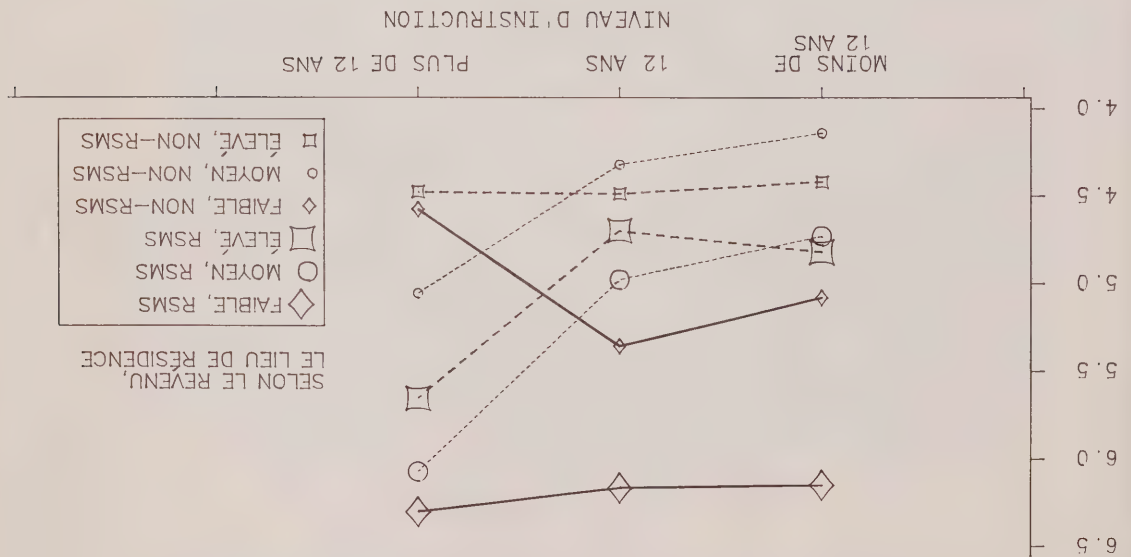


Figure 4: Moyenne prédite selon le modèle du nombre annuel de visites chez les médecins par habitant, États-Unis, 1973.

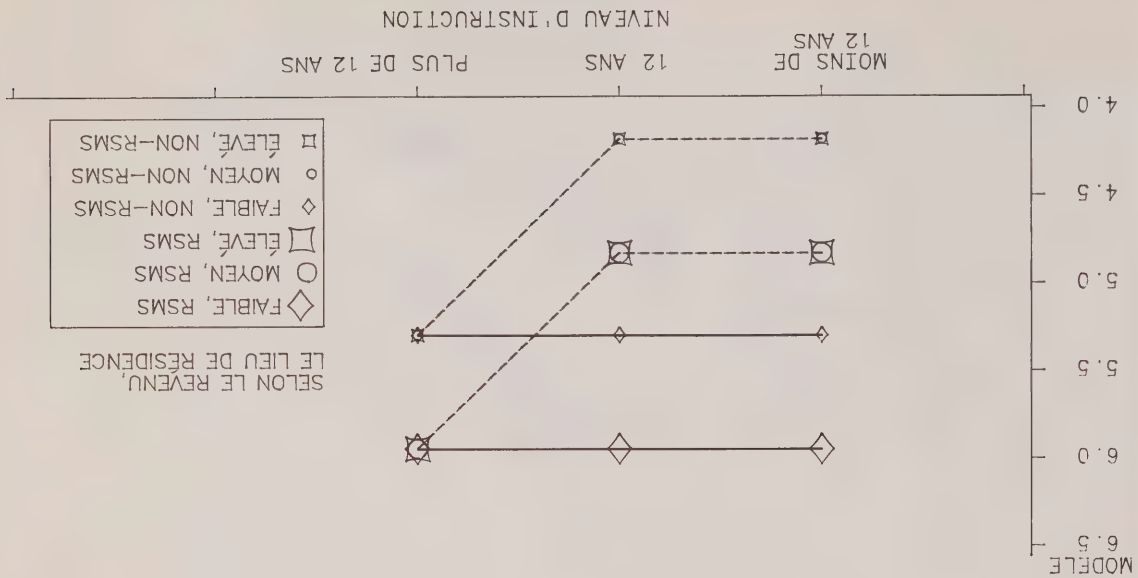


Figure 1: Moyennes observées de l'enquête sur la fécondité au Sri Lanka, 1975. Source de données: Little (1982).

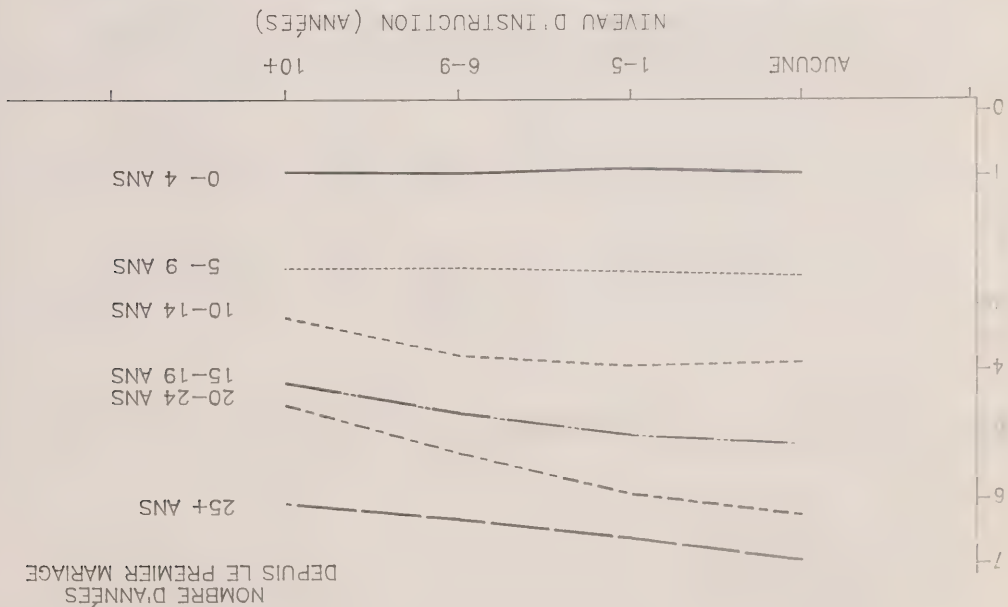


Figure 2: Moyennes ajustées de l'enquête sur la fécondité au Sri Lanka, 1975. Source de données: Little (1982).

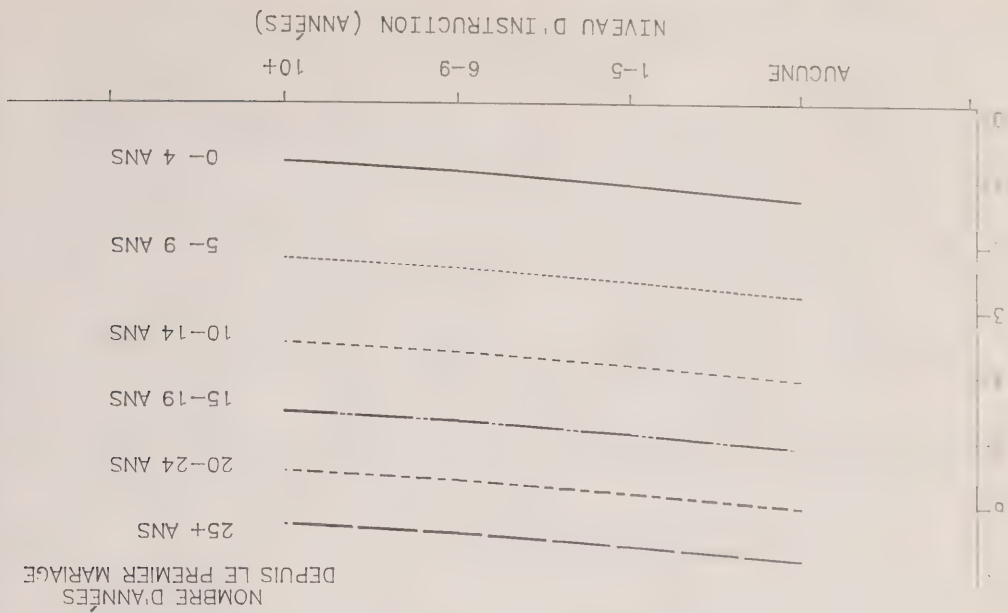




Tableau 6: Valeurs brutes et ajustées des taux de sélection après le test 2 de l'enquête canadienne sur la santé et les personnes handicapées, hommes âgés de 15 à 64 ans, selon la catégorie d'activité sur le marché du travail et l'indice de recours à un enquête-substitut, Canada, janvier 1983 (résultats non pondérés)

Age	Aucun enquête-substitut	Enquête-substitut: homme	Enquête-substitut: femme
-----	-------------------------	--------------------------	--------------------------

Actifs			
15 - 24	val. brutes val. ajustées	val. brutes val. ajustées	val. brutes val. ajustées
25 - 34	val. brutes val. ajustées	val. brutes val. ajustées	val. brutes val. ajustées
35 - 44	val. brutes val. ajustées	val. brutes val. ajustées	val. brutes val. ajustées
45 - 54	val. brutes val. ajustées	val. brutes val. ajustées	val. brutes val. ajustées
55 - 64	val. brutes val. ajustées	val. brutes val. ajustées	val. brutes val. ajustées

Inactifs			
15 - 24	val. brutes val. ajustées	val. brutes val. ajustées	val. brutes val. ajustées
25 - 34	val. brutes val. ajustées	val. brutes val. ajustées	val. brutes val. ajustées
35 - 44	val. brutes val. ajustées	val. brutes val. ajustées	val. brutes val. ajustées
45 - 54	val. brutes val. ajustées	val. brutes val. ajustées	val. brutes val. ajustées
55 - 64	val. brutes val. ajustées	val. brutes val. ajustées	val. brutes val. ajustées

NOTE: Les chiffres entre parenthèses sont les erreurs-types.

Tableau 5: Nombre estimé de visites chez un médecin, données du tableau 4 et valeurs ajustées

Niveau d'instruction (années d'études)	Revenu de la famille	RMS			NON-RMS		
		0 à 4,999	5,000 à 14,999	15,000 ou plus	Moins de 12 ans	12 ans	Plus de 12 ans
					valeur orig. 6.15 (0.18) valeur ajustée 5.95 (0.07) fcart 0.20	valeur orig. 6.17 (0.41) valeur ajustée 5.95 (0.07) fcart 0.22	valeur orig. 6.31 (0.49) valeur ajustée 5.95 (0.07) fcart 0.36
					valeur orig. 4.73 (0.13) valeur ajustée 4.83 (0.07) fcart -0.10	valeur orig. 4.98 (0.17) valeur ajustée 4.83 (0.07) fcart 0.15	valeur orig. 6.08 (0.19) valeur ajustée 5.95 (0.07) fcart 0.13
					valeur orig. 4.82 (0.25) valeur ajustée 4.83 (0.07) fcart -0.01	valeur orig. 4.70 (0.18) valeur ajustée 4.83 (0.07) fcart -0.13	valeur orig. 5.66 (0.16) valeur ajustée 5.95 (0.07) fcart -0.29
					valeur orig. 5.08 (0.26) valeur ajustée 5.30 (0.11) fcart -0.22	valeur orig. 4.14 (0.15) valeur ajustée 4.18 (0.11) fcart -0.04	valeur orig. 4.32 (0.19) valeur ajustée 4.18 (0.11) fcart 0.14
					valeur orig. 5.36 (0.44) valeur ajustée 5.30 (0.11) fcart 0.06	valeur orig. 4.32 (0.19) valeur ajustée 4.18 (0.11) fcart 0.14	valeur orig. 4.58 (0.58) valeur ajustée 5.30 (0.11) fcart -0.72
					valeur orig. 4.42 (0.37) valeur ajustée 4.18 (0.11) fcart 0.24	valeur orig. 4.49 (0.33) valeur ajustée 4.18 (0.11) fcart 0.31	valeur orig. 4.48 (0.31) valeur ajustée 5.30 (0.11) fcart -0.82
					valeur orig. 5.08 (0.26) valeur ajustée 5.30 (0.11) fcart -0.22	valeur orig. 4.14 (0.15) valeur ajustée 4.18 (0.11) fcart -0.04	valeur orig. 4.32 (0.19) valeur ajustée 4.18 (0.11) fcart 0.14
					valeur orig. 5.36 (0.44) valeur ajustée 5.30 (0.11) fcart 0.06	valeur orig. 4.32 (0.19) valeur ajustée 4.18 (0.11) fcart 0.14	valeur orig. 4.58 (0.58) valeur ajustée 5.30 (0.11) fcart -0.72
					valeur orig. 4.42 (0.37) valeur ajustée 4.18 (0.11) fcart 0.24	valeur orig. 4.49 (0.33) valeur ajustée 4.18 (0.11) fcart 0.31	valeur orig. 4.48 (0.31) valeur ajustée 5.30 (0.11) fcart -0.82

Tableau 3: Analyse de la variance des données du tableau 1

Source	Effets principaux	Durée du mariage Niveau d'instr.	Interactions	Durée du mariage x niveau d'instr. résiduelle	Total
Somme des carrés (SC)	27402.684	225.535	0.004	206.965	55565.031
Proportion du total des SC	0.493	0.004	0.004	0.499	
Degrés de liberté	5	3	15	6787	6810
Carré moyen	5480.537	75.178	13.798	4.986	
F	1340.990	18.395	3.376		
Niveau de signifi- cation	.000	.000	.000	.000	

Tableau 4: Nombre annuel de visites chez un médecin par habitant selon la taille de la localité, le revenu de la famille et le niveau d'instruction du chef de famille, États-Unis, 1973

Revenu de la famille		Niveau d'instruction (années d'études)	
		0 à 4,999	5,000 à 14,999
		15,000 ou plus	
RSMs			
Moins de 12 ans	6.15 (0.18)	4.73 (0.13)	4.82 (0.25)
12 ans	6.17 (0.41)	4.98 (0.17)	4.70 (0.18)
Plus de 12 ans	6.31 (0.49)	6.08 (0.19)	5.66 (0.16)
NON-RSMs			
Moins de 12 ans	5.08 (0.26)	4.14 (0.15)	4.42 (0.37)
12 ans	5.36 (0.44)	4.32 (0.19)	4.49 (0.33)
Plus de 12 ans	4.58 (0.58)	5.06 (0.29)	4.48 (0.31)

Note: Les chiffres entre parenthèses indiquent l'erreur-type des estimations.

Tableau 2: Interactions influant sur le nombre moyen d'enfants  
(données du tableau 1)

Niveau d'instruction (années d'études)	Nombre d'années depuis le premier mariage						
		10 et plus	6 à 9	1 à 5	0	5 à 9	10 à 14
		0.92	0.92	0.88	0.96	Moyenne brute 1.31 Moy. ajustée -0.35 Interaction -0.19	Moyenne brute 3.87 Moy. ajustée 4.06 Interaction -0.19
		2.44	2.39	2.46	2.54	Moyenne brute 2.78 Moy. ajustée -0.24 Interaction -0.08	Moyenne brute 3.91 Moy. ajustée 3.82 Interaction 0.09
		3.76	3.14	3.73	3.87	Moyenne brute 3.46 Moy. ajustée -0.32 Interaction -0.12	Moyenne brute 4.61 Moy. ajustée 4.51 Interaction -0.05
		3.77	3.46	3.61	4.06	Moyenne brute 4.82 Moy. ajustée -0.38 Interaction 0.02	Moyenne brute 5.13 Moy. ajustée 5.11 Interaction 0.02
		4.84	4.13	4.97	5.13	Moyenne brute 4.47 Moy. ajustée -0.94 Interaction 0.01	Moyenne brute 5.87 Moy. ajustée 5.77 Interaction 0.10
		5.79	4.47	5.22	6.22	Moyenne brute 5.97 Moy. ajustée -0.25 Interaction 0.10	Moyenne brute 6.92 Moy. ajustée 6.82 Interaction 0.10
		6.65	5.97	6.23	6.92	Moyenne brute 3.94 Moy. ajustée -0.14 Interaction 0.03	Moyenne brute 4.24 Moy. ajustée 3.99 Interaction 0.25
		3.94	2.30	3.26	5.17	Moyenne brute 3.63 Moy. ajustée -0.17 Interaction 0.03	Moyenne brute 4.24 Moy. ajustée 3.99 Interaction 0.25

Tableau 1: Nombre moyen d'enfants nés, selon la durée du mariage et le niveau d'instruction. Sri Lanka, 1975 (source: Little: 1982)

Nombre d'années depuis le premier mariage	Niveau d'instruction (années d'études)											
			0	1 à 5	6 à 9	10 et plus	Rangées					
n à 4	Moyenne	0.96	0.88	0.95	0.92	0.92	1281					
	Total	112	376	442	351							
5 à 9	Moyenne	2.54	2.46	2.39	2.39	2.44	1231					
	Total	172	442	362	255							
10 à 14	Moyenne	3.87	3.91	3.73	3.14	3.76	1117					
	Total	197	482	293	145							
15 à 19	Moyenne	5.13	4.97	4.61	4.13	4.84	1057					
	Total	239	461	262	95							
20 à 24	Moyenne	6.22	5.87	5.22	4.47	5.79	893					
	Total	292	377	184	40							
25 et plus	Moyenne	6.92	6.55	6.23	5.97	6.65	1232					
	Total	501	548	161	22							
Colomes	Moyenne	5.17	4.24	3.26	2.30	3.94	6811					
	Total	1513	2686	1704	908							

- [2] Hocking, R.R. (1983). Developments in linear regression methodology: 1959-1982 (comprend une analyse). Technometrics 25, pp. 219-249.
3. Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Statist. 1, pp. 799-821.
- [4] Koch, G.G., Gillings, D.R. et Stokes, M.E. (1980). Biostatistical implications of design, sampling and measurement to health science data analysis. Ann. Rev. Public Health 1, pp. 163-225.
- [5] Landwehr, J.M., Pregibon, D. et Shoemaker, A. (1984). Graphical methods for assessing logistic regression models (comprend une analyse). J. Amer. Statist. Assoc. 79, pp. 61-83.
- [6] Little, R.J.A. (1982). Direct standardization: A tool for teaching linear models for unbalanced data. Amer. Statist. 36, pp. 38-43.
- Nelder, J.A. et Wedderburn, R.W.M. (1972). Generalized linear models. J. R. Statist. Soc. A, 135, pp. 370-384.



en faisant les définitions suivantes:

$$\begin{aligned} \hat{w}_j(t) &= \hat{u}_j(t) \\ \hat{z}_j(t) &= \log \hat{u}_j(t) + \frac{\hat{u}_j(t)}{\gamma_j - \hat{u}_j(t)} \end{aligned}$$

Ainsi, des modèles semblables à ceux décrits à la section 3 peuvent être analysés de manière analogue dans le cadre d'un modèle linéaire généralisé.

## 5. DIAGNOSTICS

Les méthodes de régression linéaire sont maintenant connues depuis plus d'un siècle (voir Hocking (1983) pour un résumé des recherches faites au cours des vingt-cinq dernières années). Les travaux x faits ces dernières années ont surtout porté sur les difficultés qui se posent lorsqu'il y a des multicolli-néarités dans les variables (qui entraînent des variances élevées dans les estimations des paramètres) et lorsqu'un modèle est inadéquat. Quelques diag-nostics qui permettent de déceler ces deux genres de problèmes sont maintenant offerts dans les logiciels SAS et SPSS-X.

Les méthodes décrites plus haut étendent la régression linéaire à une grande variété de problèmes. Les nouveaux diagnostics applicables à ce genre de modèle sont présentés dans un article de Landwehr, Pregibon et Shoemaker (1984).

Dans beaucoup d'analyses statistiques, le modèle proposé n'est qu'une approximation de la réalité. Par conséquent, l'utilisateur de ces modèles doit prendre en compte ces outils diagnostiques au cours de son analyse.

## BIBLIOGRAPHIE

- [1] Dolson, D. et Morin, J.-P. (1983). Disability data development project: Analysis of screening questionnaires. Document technique de la Division de la santé, Statistique Canada.

répondre itérativement", En particulier, les poids à la t<sup>ième</sup> itération ont la valeur

$$w_j(t) = \frac{1}{\sum_j w_j(t) [q_j(t)]^2}$$

on obtient la valeur des variables dépendantes à la t<sup>ième</sup> itération à partir de l'équation

$$\hat{z}_j(t) = q_j(t) + q_j(t) (y_j - \hat{u}_j(t)).$$

Le résultat de la (t + 1)<sup>ième</sup> itération de  $\hat{\beta}$  est donc la solution de l'équation

$$\sum_j w_j(t) [\hat{z}_j(t) - \sum_j x_{kj} \hat{\beta}_{kj}(t+1)] x_{kj} = 0.$$

La matrice des covariances estimées de  $\hat{\beta}$  est  $A^{-1}$ , où le (k, l)<sup>ième</sup> élément de la matrice A est

$$a_{kl} = \sum_j w_j x_{kj} x_{lj}.$$

On peut donc utiliser les programmes habituels de régression par la méthode des moindres carrés pour exécuter ces modèles linéaires généralisés.

Par exemple, dans la méthode très répandue d'analyse des tableaux de contingence à partir de modèles dits "log-linéaires", le point de départ est un modèle fondé sur la loi de Poisson, où  $\log u_j = \sum_i x_{ij} \beta_i$ . Dans ce modèle,

$$V_j = u_j,$$

$$q(u_j) = \log u_j,$$

on obtient la solution des moindres carrés pondérés itérativement

Par ailleurs, il n'y a pas d'interaction entre l'indice de recours à un enquête-substitut et les variables âge et catégorie d'activité. Les résultats n'indiquent pas nécessairement la présence systématique d'un biais attribuable au recours à un enquête-substitut, mais ils révèlent qu'un tel biais est possible. Sans une étude spéciale comme, par exemple, un programme de réinterview des enquêtes-substituts, il est impossible de conclure avec certitude que ce biais existe.

## 4.2 Modèles linéaires généralisés

Dans la section précédente, nous avons décrit un éventail de modèles linéaires fondés sur la loi binomiale et dont l'analyse des probits et la régression logistique constituent des cas spéciaux. Nous étendons maintenant ces modèles à la famille exponentielle selon l'analyse proposée par Nelder et Wedderburn (1972).

Comme à la section 2.3, nous supposons que  $Y_j$  a une distribution exprimée par la fonction

$$f(y_j) = \exp[k_j y_j \theta_j - b(\theta_j)] + c(y_j, k_j),$$

où  $\mu_j = E[Y_j] = b'(\theta_j)$  et  $V_j = \text{Var}[Y_j] = b''(\theta_j)/k_j$ .

Nous définissons  $\eta_j = q(\mu_j) = \sum_{i=1}^I X_{ij} \beta_i$  comme étant la composante linéaire du modèle, où  $q(\cdot)$  est une fonction connue.

Pour obtenir les estimations du maximum de vraisemblance de  $\beta$ , il faut résoudre le système d'équations

$$\sum_j \frac{(Y_j - \mu_j) X_{ij}}{V_j [q'(\mu_j)]} = 0.$$

Nelder et Wedderburn (1972) ont démontré qu'un moyen raisonnable d'estimer  $\beta$  est d'effectuer une série de régressions par la méthode des moindres carrés pondérés, où les valeurs des poids et des variables dépendantes sont révisées après chaque itération. Cette technique a été baptisée "méthode des moindres

sont illustrées à la figure 5. Le modèle ajusté a permis de réduire le nombre de paramètres de trente à onze. Le modèle final était défini par l'équation

$$\log[p_{ijk}/(1 - p_{ijk})] = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ijk},$$

où  $\sum \alpha_i = \sum \beta_j = \sum \gamma_k = 0$ ,  $\sum \delta_{ijk} = 0$ ,  $\sum \delta_{ij.} = 0$ ,  $\sum \delta_{.j.} = 0$ ,  $\sum \delta_{.k.} = 0$ , pour la  $i^{\text{ème}}$  catégorie d'âge, la  $j^{\text{ème}}$  catégorie d'activité sur le marché du travail et le  $k^{\text{ème}}$  indice de recours à un enquête-substitut (deux possibilités, selon qu'il a y eu ou qu'il n'y a eu pas d'enquête-substitut). Les valeurs estimées des paramètres sont

Modèles	Estimation
$\mu$	-1.43
$\alpha$	15 à 24 ans -1.12 25 à 34 ans -0.571 35 à 44 ans 0.0143 45 à 54 ans 0.629 55 à 64 ans 1.05
$\beta$	Actifs -0.576 Inactifs 0.576
$\gamma$	- enquête-substitut 0.0859 + enquête-substitut -0.0859
$\delta$	actifs de 15 à 24 ans 0.385 actifs de 25 à 34 ans 0.0938 actifs de 35 à 44 ans -0.175 actifs de 45 à 54 ans -0.243 actifs de 55 à 64 ans -0.0612 inactifs de 15 à 24 ans -0.385 inactifs de 25 à 34 ans -0.0938 inactifs de 35 à 44 ans 0.175 inactifs de 45 à 54 ans 0.243 inactifs de 55 à 64 ans 0.0612

Les valeurs ajustées sont illustrées à la figure 6. On peut voir que, même quand on tient compte de l'âge et de la catégorie d'activité, le recours à un enquête-substitut a un effet sur les taux de sélection. Cet effet ne semble pas dépendre du sexe de l'enquête-substitut.

de sorte que l'estimation du paramètre recherché est contenue dans la solution de l'équation

$$\sum_{i=1}^J (y_i - \hat{p}_j) x_{ij} = 0, \quad \text{pour } i = 0, \dots, r.$$

La matrice des covariances de  $\hat{\beta}_0, \dots, \hat{\beta}_r$  est  $A^{-1}$ , où  $A$  est une matrice dont le  $(k, \lambda)$  ième élément est

$$A_{k\lambda} = \frac{\sum_{i=1}^J x_{ki} x_{\lambda i}}{\sum_{i=1}^J p_i (1 - p_i) \{q'(p_i)\}^2}$$

Ce résultat peut être utilisé pour établir des intervalles de confiance, faire des tests d'hypothèses et construire des modèles.

Dans la régression logistique, la covariance se ramène à l'expression

$$A_{k\lambda} = \sum_{j=1}^J p_j (1 - p_j) x_{kj} x_{\lambda j}.$$

Pour illustrer l'utilité de ces modèles, nous examinerons une analyse non publiée faite par Dolson et Morin à partir de données de l'enquête canadienne sur la santé et l'invalidité. La variable dépendante indiquait si une personne était classée comme handicapée après l'exécution du test de sélection 2 de l'enquête supplémentaire sur les personnes handicapées qui a été menée en même temps que l'enquête sur la population active de janvier 1983. (Pour une description détaillée de cette enquête, voir Dolson et Morin: 1983) L'analyse a été limitée aux hommes âgés de 15 à 64 ans. Parmi les 13,897 répondants, 14.4 % (résultat non pondéré) ont été sélectionnés. Comme on peut le voir au tableau 6, le taux de sélection a été ventilé en fonction du groupe d'âge, de la catégorie d'activité sur le marché du travail (deux catégories: actif et inactif) et d'une variable indiquant si un enquête-substitut a répondu à la place d'une autre personne (trois catégories ont été établies: il n'y a pas eu d'enquête-substitut, l'enquête-substitut était un homme ou l'enquête-substitut était une femme). (Le tableau 6 contient également les valeurs ajustées à partir du modèle décrit plus bas.) Ces données

Deux autres techniques très répandues sont l'analyse des probits et la régression logistique. Dans l'analyse des probits, on suppose que  $\eta_j = \Phi(\sum_{i=1}^I X_{ij} \beta_i)$ , où  $\Phi$  est la fonction cumulative d'une variable aléatoire normale standardisée. Dans la régression logistique, on suppose que

$$\eta_j = \log[p_j / (1 - p_j)] = \sum_{i=1}^I X_{ij} \beta_i.$$

Ces deux méthodes sont des outils analytiques importants, et on les retrouve dans un grand nombre de logiciels statistiques (tels que SAS et BMDP). Pour avoir une idée générale de ces deux approches, définissons

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^I X_{ij} \hat{\beta}_i.$$

Dans l'analyse des probits,  $\eta_j = \Phi^{-1}(p_j)$ , alors que dans la régression logistique,  $\eta_j = \log[p_j / (1 - p_j)]$ . Pour calculer l'estimateur du maximum de vraisemblance de  $\beta_0, \dots, \beta_I$ , il faut résoudre l'équation

$$\sum_{j=1}^J \frac{X_{ij} (p_j - \hat{p}_j)}{\hat{p}_j (1 - \hat{p}_j)} = 0, \text{ pour } i = 0, \dots, I,$$

où  $q(p_j) = \sum_{i=1}^I X_{ij} \beta_i$ . Ce système doit souvent être résolu par itération. Pour l'analyse des probits, on peut écrire

$$q'(p_j) = \frac{1}{1 - p_j} \phi(q(p_j))$$

où  $\phi(\cdot)$  est la fonction de densité de la loi normale réduite. Dans le cas de la régression logistique,

$$q'(p_j) = \frac{1}{p_j(1 - p_j)}.$$



On peut constater que l'ajustement du modèle est assez bon. Nous avons réduit des données qui comprenaient dix-huit paramètres à trois résumés statistiques et nous obtenons également des erreurs-types moins élevées qu'auparavant.

#### 4. MODÈLES LINÉAIRES GÉNÉRALISÉS

##### 4.1 Régression avec une variable dépendante dichotomique

Un des problèmes qui se posent souvent dans l'utilisation des modèles linéaires décrits à la section 3 est que les erreurs sont supposées suivre une loi normale. Il est vrai que des analyses semblables à celles abordées à la section 3 sont faisables même si les erreurs n'ont pas une distribution normale, mais seulement si la variance des erreurs satisfait à la condition  $\sigma_j^2 = \sigma^2/w_j$  et s'il n'y a pas de corrélation entre les erreurs. Dans un tel cas, les estimateurs linéaires que nous avons décrits produisent des estimations centrées à variance minimale pour les paramètres du modèle, mais il existe peut-être de meilleurs estimateurs (c'est-à-dire des estimateurs non linéaires). La recherche de meilleurs estimateurs a mené à la conception de modèles linéaires généralisés (voir Nelder et Wedderburn: 1972) et d'estimateurs robustes (voir Huber: 1973).

Par exemple, supposons que la variable dépendante  $y_j$  peut prendre seulement deux valeurs, 0 ou 1. Nous voulons construire un modèle pour exprimer  $P(y_j = 1)$  en fonction de l'expression linéaire  $X_{0j}\beta_0 + X_{1j}\beta_1 + \dots + X_{rj}\beta_r$ . Trois méthodes sont généralement utilisées pour résoudre ce problème. La première approche est de calculer  $\beta_0, \dots, \beta_r$  à l'aide de l'estimateur habituellement employé dans un modèle de régression classique. Cette technique est analogue à une analyse discriminante dans laquelle les variables  $X_{0j}, \dots, X_{rj}$  sont considérées non comme des constantes fixes et connues, mais comme des variables aléatoires (qui suivent une loi normale à plusieurs caractères et ont une matrice des covariances constante) dont la moyenne dépend de la valeur de  $y_j$ . Le principal inconvénient de cette méthode est que  $\hat{y}_j = X_{0j}\hat{\beta}_0 + \dots + X_{rj}\hat{\beta}_r$  ne peut pas être utilisé directement pour prédire la valeur de  $p_j$ . En outre, les  $X_{1j}$  sont souvent des variables qualitatives (telles que la profession, la profession, etc.), ce qui annule l'hypothèse de normalité à plusieurs caractères.

où A est une matrice dont le  $(k, l)$ ème élément est  $\sum_j w_j x_{kj} x_{lj}$ . Pour estimer  $\sigma^2$ , nous utilisons la formule  $\hat{\sigma}^2 = \frac{1}{2} w_i^T (y_i - \hat{y}_i)^2 / (n - r - 1)$ .

Le manuel nombre de logiciels comprennent des commandes faciles pour faire des tests d'hypothèses sur  $\beta$  à partir de la matrice des covariances estimées  $\hat{\sigma}^2 A^{-1}$  et des valeurs critiques de la distribution appropriée de la variable F (comme, par exemple, PROC ANOVA et PROC GLM dans le logiciel SAS).

On peut illustrer ce genre de modèle à l'aide des données présentées au tableau 4, qui est tiré d'une étude de Koch, Gillings et Stokes (1980) et indique le nombre annuel de visites chez les médecins par habitant, aux États-Unis, en 1973 selon la taille de la ville (deux catégories: 1. RSM5 = région métropolitaine standard (Standard Metropolitan Statistical Area) et 2. Non-RSM5) et le niveau d'instruction (trois catégories). Ces données proviennent de la Health Interview Survey de 1973, enquête basée sur un échantillon probabiliste complexe. La figure 3 offre une représentation graphique de ces données.

Un modèle de régression et quelques tests statistiques ont permis d'obtenir un modèle réduit qui a la forme suivante:

$$\hat{y}_i = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j},$$

où  $x_{1j} = 1$  si la  $j$ ème personne habite une RSM5

= 0 autrement

$x_{2j} = 1$  si le revenu de la famille de la  $j$ ème personne était inférieur à \$5,000 ou si la  $j$ ème personne avait plus de douze années d'instruction, dans le cas d'un chef de famille.

= 0 autrement.

Les paramètres estimés étaient  $\hat{\beta}_0 = 4.18$  (erreur-type de 0.11),  $\hat{\beta}_1 = 0.65$  (erreur-type de 0.11) et  $\hat{\beta}_2 = 1.12$  (erreur-type de 0.09). Les erreurs-types calculées ici ne sont pas celles décrites plus haut parce que les auteurs ont utilisé la matrice de dimension  $18 \times 18$  qui contient les covariances estimées de l'enquête susmentionnée, pour obtenir les erreurs-types. Cette méthode ne suit pas l'analyse de données d'enquêtes complexes.

Les résultats sont résumés au tableau 5 et sont illustrés à la figure 4.

où  $X_{0j}$ ,  $X_{1j}$ , ...,  $X_{rj}$  sont des constantes connues et  $\beta_0$ ,  $\beta_1$ , ...,  $\beta_r$  sont des coefficients inconnus. Nous supposons que les  $\varepsilon$  sont indépendants, suivent une loi normale et ont pour variance  $\sigma_j^2 = \sigma^2/w_j$ , où les  $w$  sont des poids connus. Par exemple, dans une analyse de la variance à un facteur, on pourrait utiliser la notation suivante.

$X_{0j} = 1$  pour toutes les valeurs de  $j$   
 $X_{1j} = 1$  si la  $j^{\text{ième}}$  unité appartient à la  $i^{\text{ième}}$  sous-population  
 $= -a_i/a_j$  si la  $j^{\text{ième}}$  unité appartient à la  $i^{\text{ième}}$  sous-population  
 $= 0$  autrement

pour  $i = 1, \dots, I - 1$ , où  $a_i$  est la somme des poids pour toutes les unités de la  $i^{\text{ième}}$  sous-population. Dans ce modèle.

$$\mu_i = \beta_0 + \beta_i \quad \text{pour } i = 1, \dots, I - 1.$$

$$\mu = \beta_0 - (a_1\beta_1 + \dots + a_{I-1}\beta_{I-1})/a_I.$$

Ainsi,  $\mu = \beta_0$  et  $\alpha_i = \beta_i$  pour  $i = 1, \dots, I - 1$ .

Une formule de régression semblable peut également être définie pour les analyses à deux facteurs et plus.

Dans la version générale du modèle de régression, les estimateurs de  $\beta_0$ , ...,  $\beta_r$  sont  $\hat{\beta}_0$ , ...,  $\hat{\beta}_r$ , les solutions de l'expression

$$\sum w_j (y_j - \hat{y}_j) X_{ij} = 0, \quad i = 0, 1, \dots, r$$

$$\text{où } \hat{y}_j = \hat{\beta}_0 X_{0j} + \hat{\beta}_1 X_{1j} + \dots + \hat{\beta}_r X_{rj}.$$

Si on veut vérifier des hypothèses, construire des modèles et établir des intervalles de confiance pour les  $\beta$ , il faut calculer la matrice des covariances des  $\hat{\beta}$ . On doit résoudre l'équation

$$\text{Var}(\hat{\beta}) = \sigma^2 A^{-1}$$

$$\mu = \bar{y} \dots$$

$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y} \dots - \frac{1}{2} \sum_k w_{ijk} \hat{\beta}_{j/k} \dots - \frac{1}{2} \sum_k w_{ijk} \hat{\alpha}_{1/k}$$

$$\hat{\alpha}_{1j} = \bar{y}_{.j.} - \bar{y} \dots - \frac{1}{2} \sum_k w_{1jk} \hat{\alpha}_{1/k} \dots - \frac{1}{2} \sum_k w_{1jk} \hat{\beta}_{j/k}$$

$$1 = \bar{y}_{11.} - \bar{y} \dots - \hat{\alpha}_1 - \hat{\beta}_j$$

où  $\bar{y}_{11.}$ ,  $\bar{y}_{1.}$ , etc. sont les moyennes pondérées appropriées.

Or, dans le modèle additif,  $\mu_{1j} = \mu + \alpha_1 + \beta_j$ . Nous avons tracé à la

figure 1 les courbes formées par les moyennes dans chaque case du tableau 1. Selon le modèle additif, toutes ces courbes doivent être parallèles. Si on

ajuste le modèle additif aux données du tableau 1, on obtient les moyennes ajustées qui figurent au tableau 2. Comme nous pouvons le constater, l'effet

de l'instruction est considérablement affaibli après l'ajustement de ce modèle. La raison est que les femmes les plus instruites n'étaient pas ma-

riées aussi longtemps que les autres, de sorte que c'est le nombre d'années depuis le premier mariage qui devient le facteur important. Toutefois, comme

le révèle l'analyse de la variance résumée au tableau 3, tous les effets prin-

cipaux et les interactions ont un caractère significatif. Le modèle additif doit donc être rejeté. Mais seulement 0.4% de la variation totale est expli-

quée par les interactions entre le niveau d'instruction et la durée du mariage, tandis que 49.7% de cette variation est expliquée par le modèle addi-

tif. Nous pouvons conclure que le modèle additif nous a permis de mieux com-

prendre les données et que l'effet du niveau d'instruction n'est pas aussi

important qu'il le semblait à l'origine.

### 3.3 Modèle de régression

Les modèles d'analyse de la variance décrits plus haut peuvent être consi-

dérés comme des cas spéciaux du modèle de régression linéaire multiple défini

par l'équation

$$y_j = \beta_0 x_{0j} + \beta_1 x_{1j} + \dots + \beta_r x_{rj} + \epsilon_j,$$

dans chaque case indique le nombre d'enfants mis au monde selon la durée du mariage et le niveau d'instruction.

Les moyennes des colonnes et des rangées semblent indiquer que le nombre moyen d'enfants augmente en fonction de la durée du mariage et diminue en fonction du niveau d'instruction. Le modèle d'analyse de la variance à deux facteurs peut être exprimé sous la forme suivante:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

où les valeurs de  $\epsilon$  sont considérées comme étant indépendantes, distribuées selon une loi normale et possédant une variance égale à  $\sigma^2_{ijk} = \sigma^2/w_{ijk}$ . Les  $w$  sont des poids connus. Dans la plupart des analyses, ces poids sont constants. Pour estimer les paramètres de ce modèle, il est nécessaire de fixer des contraintes à ces paramètres afin d'éviter qu'ils ne soient pas uniques. Les conditions habituellement ajoutées au modèle sont:

$$\sum_i \sum_j \sum_k w_{ijk} \alpha_i = 0,$$

$$\sum_i \sum_j \sum_k w_{ijk} \beta_j = 0,$$

$$\sum_i \sum_j \sum_k w_{ijk} \gamma_{ij} = 0,$$

$$\sum_j \sum_k w_{ijk} \gamma_{ij} = 0.$$

Les estimateurs sont définis par le système d'équations

$$\frac{\partial \mu_{ij}}{\partial \theta} = 0$$

où  $\hat{\theta}_1, \hat{\theta}_2, \dots$  correspondent aux estimations des paramètres  $\mu, \alpha_i, \beta_j, \gamma_{ij}$ , etc. Les valeurs des  $\alpha$  et des  $\beta$  représentent les effets principaux, tandis que celles des  $\gamma$  sont les interactions entre les deux facteurs. On obtient ainsi les estimateurs suivants.

## 3.2 Analyse de la variance à deux facteurs

L'extension de cette représentation est particulièrement utile pour les modèles d'analyse de la variance à deux facteurs ou plus, ce qui fait l'objet des sections 3.2 et 3.3. Un des problèmes les plus importants dans ce genre de modèle est de vérifier si toutes les moyennes sont égales. Autrement dit, on veut savoir si  $\mu_1 = \mu_2 = \dots = \mu_I$  ou  $\alpha_1 = \alpha_2 = \dots = \alpha_I = 0$ . Il existe des logiciels standard (SAS, SPSS, etc.) qui comprennent ces tests d'hypothèses dans leurs sous-programmes d'analyse de la variance (ANOVA). Une question connexe est: quelles sous-populations ont une moyenne égale s'il s'avère que les moyennes ne sont pas toutes égales? Quand on ne peut plus ajouter d'autres facteurs à l'analyse comme, par exemple, dans une analyse de la variance à deux facteurs, il faut alors procéder à une comparaison multiple de moyennes. Un grand nombre de logiciels statistiques permettent d'exécuter ce type d'analyse.

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

$$0 \leq \sum w_{ij} \alpha_i = 0 \text{ et } \mu = \sum w_{ij} \mu_i / \sum w_{ij}$$

$$\alpha_i = \mu_i - \mu.$$

Une manière différente mais équivalente de formuler ce modèle est la suivante. Soient les hypothèses de ce modèle, les moyennes estimées sont indépendantes, obéissent à une loi normale et  $E(\hat{\mu}_i) = \mu_i$  et  $\text{Var}(\hat{\mu}_i) = \sigma^2 / \sum w_{ij}$ . Ces propriétés permettent d'établir des intervalles de confiance pour les moyennes

$$\hat{\mu}_i = \sum w_{ij} Y_{ij} / \sum w_{ij}.$$

Les données présentées au tableau 1 sont tirées de l'enquête sur la fécondité menée en 1975 au Sri Lanka (voir Little: 1982). La moyenne qui figure



$$\sum w_j [y_j \theta - b(\theta)] + \frac{\partial c(y_j, k_j)}{\partial k_j} = 0.$$

L'autre moyen d'estimer  $\hat{V}(u)$  sans biais est d'utiliser l'estimateur suivant qui est lié moins étroitement à un modèle :

$$\hat{V}_1(u) = \frac{\sum w_j^2 (y_j - u)^2}{(n-1) \sum w_j^2}.$$

Cet estimateur offre une autre façon de construire les intervalles de confiance de  $\mu$ . La principale hypothèse sur laquelle repose la validité de cette méthode est que  $\text{Var}(y_j) \propto 1/w_j$ .

### 3. MODÈLES LINÉAIRES

#### 3.1 Analyse de la variance à un facteur

Le modèle de l'analyse de la variance à un facteur constitue une extension simple des modèles à une variable basés sur la loi normale qui sont décrits à la section 2.2. Dans l'analyse de la variance, nous observons une caractéristique de chaque unité échantillonnée, mais nous définissons également des sous-populations. Ces sous-populations peuvent être des groupes d'âge et être représentée par l'expression

$$y_{ij} = \mu_i + \varepsilon_{ij}; i = 1, \dots, I; j = 1, \dots, n_i,$$

où les  $\mu$  sont les moyennes réelles qui varient d'une sous-population à une autre, et, par hypothèse, les  $\varepsilon$  sont indépendants et suivent une loi normale avec une variance égale à  $\sigma^2/w_{ij}$ , où les  $w_{ij}$  sont des poids connus. Dans la plupart des analyses, ces poids sont constants. L'estimateur habituel de  $\mu_i$  dans ce modèle est

$$E(y_j) = \sigma^2, \quad \text{Var}(y_j) = 2\sigma^4/v_j,$$

$$\theta = -1/\sigma^2,$$

$$k_j = v_j/2,$$

$$b(\theta) = -\log(-\theta).$$

Comme on peut le noter dans ces exemples, la famille exponentielle comprend un grand nombre de distributions fréquentes. En général:

$$E(y_j) = b'(\theta) = \mu, \quad \text{Var}(y_j) = b''(\theta)/k_j = v_j$$

où  $b'(\cdot)$  et  $b''(\cdot)$  représentent la dérivée première et la dérivée seconde de  $b(\cdot)$ .

Si les valeurs de  $y_1, \dots, y_n$  sont indépendantes, l'estimation du maximum de vraisemblance de  $\theta$  provient de la solution de:

$$\mu = \sum k_j y_j / \sum k_j = \sum w_j y_j / \sum w_j$$

où  $\mu = b'(\theta)$ . Il doit donc exister une grande famille de modèles où une moyenne pondérée d'un échantillon constitue un estimateur efficace de la moyenne de la population. L'estimateur de la variance de  $\mu$  est

$$\hat{\text{Var}}(\hat{\mu}) = (\sum k_j^2 v_j) / (\sum k_j)^2$$

$$= b''(\theta) / (\sum k_j).$$

Pour les grands échantillons, l'intervalle de confiance de 95 % de  $\mu$  est  $\hat{\mu} \pm 1.96 [\hat{\text{Var}}(\hat{\mu})]^{1/2}$ , pourvu que le modèle soit vrai.

Dans les cas où  $k_j = kw_j$  est connu seulement jusqu'à la valeur de la cons-  
tante de proportionnalité,  $k$ , (comme, par exemple, dans un modèle fondé sur la loi normale), il est nécessaire d'estimer la valeur de  $k$ . L'estimation du maximum de vraisemblance dépend de la solution à l'équation:

Exemple 2 (Variable normale)

Soit  $y_j$  une variable qui suit une loi normale et qui a pour moyenne  $\mu$  et pour variance  $\sigma_j^2$ . Ainsi, on peut écrire :

$$f(y_j) = \frac{1}{\sigma_j} \exp \left\{ -\frac{1}{2} \left( \frac{y_j - \mu}{\sigma_j} \right)^2 \right\} ; -\infty < y_j < \infty$$

$$E(y_j) = \mu, \quad \text{Var}(y_j) = \sigma_j^2,$$

$$\theta = \mu,$$

$$k_j = 1/\sigma_j^2,$$

$$b(\theta) = \mu^2/2.$$

Exemple 3 (Moyenne d'une variable de Poisson)

Soit  $y_j$  une variable de Poisson qui a pour moyenne  $n_j \lambda$ . Définissons ensuite  $\bar{y}_j = y_j/n_j$ . On a alors :

$$f(\bar{y}_j) = e^{-n_j \lambda} \frac{(n_j \lambda)^{n_j \bar{y}_j}}{n_j!};$$

$$\bar{y}_j = 0, \frac{1}{n_j}, \frac{2}{n_j}, \dots,$$

$$E(\bar{y}_j) = \lambda, \quad \text{Var}(\bar{y}_j) = \lambda/n_j$$

$$\theta = \log \lambda,$$

$$k_j = n_j,$$

$$b(\theta) = e^\theta.$$

Exemple 4 ( $\chi^2$ )

Soit  $y_j$  une variable qui suit une loi de type  $\chi^2_{\nu_j}/\nu_j$ . Ce genre de variable est souvent utilisée dans les modèles d'analyse et de décomposition de la variance, où  $y_j$  est le carré moyen. On a alors :

$$f(y_j) = \frac{1}{\nu_j} \left( \frac{\nu_j}{2} \right)^{\nu_j/2} \exp \left\{ -\frac{\nu_j y_j}{2} \right\} / \Gamma(\nu_j/2); \quad y_j \geq 0,$$

avec les poids d'échantillonnage utilisés dans les plans de sondage complexes élaborés pour des populations finies. Quand un analyste ajuste des données recueillies à partir d'un plan de sondage complexe conçu pour une population finie, il peut décider d'inclure à la fois des poids théoriques et des poids d'échantillonnage dans le calcul d'une estimation.

## 2.3 Modèles appartenant à la famille exponentielle

Dans les modèles décrits plus haut, les lois binomiale et normale peuvent être considérées comme des cas spéciaux d'un ensemble beaucoup plus général de distributions connu sous le nom de "famille exponentielle". La forme générale que nous utiliserons pour exprimer cette distribution est :

$$f(y_j) = \exp \{ k_j^T y_j - b(\theta) \} + c(y_j, k_j^T) 1,$$

où les valeurs de  $y_j$  ne dépendent pas de  $\theta$ .

Nous supposons que  $k_j = k w_j$ , où les poids  $w_1, \dots, w_n$  sont connus. Dans un grand nombre de cas, la valeur de  $k$  est également connue.

### Exemple 1 (proportion binomiale)

Soit  $\bar{y}_j = y_j/n_j$  une proportion calculée pour un échantillon de  $n_j$  observations d'une population binomiale. On peut donc écrire :

$$y_j = n_j \bar{y}_j \quad (1 - \bar{y}_j)^{n_j} \bar{y}_j^{n_j} (1 - p)^{n_j} (1 - p)^{n_j} : \bar{y}_j = 0, \frac{1}{2}, \frac{n_j}{2}, \dots, 1,$$

$$E(\bar{y}_j) = p, \text{Var}(\bar{y}_j) = p(1 - p)/n_j,$$

$$\theta = \log[p/(1 - p)],$$

$$k_j = n_j,$$

$$b(\theta) = \log(1 + e^\theta).$$

$$f(y) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} (y - \mu)^2\right\}; -\infty < y < \infty.$$

La moyenne de la population est  $\mu$  et c'est souvent ce paramètre qu'on veut estimer. La variance de la population est  $\sigma^2$ .

Si nous observons des données  $y_1, y_2, \dots, y_n$  dans la population, l'estimateur habituel de  $\mu$  est  $\bar{y} = \sum y_i / n$ . L'estimateur de l'erreur-type de  $\mu$  est  $s(\mu) = s / \sqrt{n}$ , où

$$s^2 = \sum (y_i - \bar{y})^2 / (n - 1).$$

Comme dans le cas de la loi binomiale, l'intervalle de confiance de 95 % est  $\bar{y} \pm 1.96s(\mu)$  pour les grands échantillons. Cet intervalle est aléatoire et a une probabilité de 95 % de couvrir la vraie valeur de  $\mu$ . Quand un échantillon est petit ( $n < 60$ ), on peut remplacer la valeur 1.96 par la valeur appropriée de la distribution du  $t$  de Student pour accroître l'exactitude des intervalles estimés. Pour obtenir différents niveaux de confiance, il suffit de substituer à la valeur 1.96 le chiffre qui correspond au percentile approprié de la loi normale réduite ou de la distribution du  $t$  de Student.

Dans certains cas, il n'est pas réaliste de supposer que la variance est constante, surtout dans les modèles linéaires décrits à la section 3. Une manière simple d'étendre l'application de la loi normale est de représenter la variance de  $X_i$  par  $\sigma_i^2$ , où  $\sigma_i^2 = \sigma^2 / w_i$ . Nous supposons ici que  $w_1, w_2, \dots, w_n$  sont des poids connus. L'estimateur de la moyenne est alors  $\mu = \sum w_i y_i / \sum w_i$ , une moyenne pondérée des observations. L'erreur-type de cette moyenne est  $s(\mu) = s(\sum w_i y_i)^{\frac{1}{2}} / \sum w_i$ , où

$$s^2 = \sum w_i (y_i - \hat{\mu})^2 / (n - 1).$$

L'estimation d'intervalles de confiance pour  $\mu$  est analogue au cas simple. Il est important de souligner que les poids  $w_1, \dots, w_n$  sont liés à la spécification des modèles basés sur la loi normale et n'ont habituellement aucun

Les deux distributions importantes utilisées dans l'élaboration de modèles pour expliquer des données est la loi normale qui est exprimée par la fonction

## 2.2 Modèles fondés sur la loi normale

consommation de tabac, le poids, etc. être liées à d'autres facteurs tels que l'âge, le sexe, l'état de santé, la pour tous les individus par un schéma dans lequel les probabilités peuvent le schéma permet de remplacer l'hypothèse de probabilités constantes. Il est important de souligner que les modèles linéaires généralisés décrits décider peuvent être vérifiées à l'aide de la loi binomiale.

éventuellement indépendants, les hypothèses sur l'invariabilité de la probabilité de ce type d'hypothèse dans leurs calculs.) Si chaque décès individuel est un pas identiques. (Les compagnies d'assurance-vie et leurs actuaires utilisent de décès au cours d'autres années, bien que les populations mères ne soient peut utiliser le nombre de décès au cours d'une année pour estimer le nombre probabilité de décès est constante sur une période de quelques années, on et tous les décès sont essentiellement des événements indépendants. Si la loi binomiale dans laquelle chaque individu a une probabilité égale de mourir peut être conçu comme étant un résultat particulier provenant d'une distribution le nombre de décès au Canada dans un groupe d'âge et sexe au cours d'une année qui, à notre avis, présentent des caractéristiques semblables. Par exemple, dans bien des cas, nous voulons faire des déductions sur d'autres populations inférence basée sur l'échantillon s'étend à la population mère. Toutefois, aléatoire simple prélevé dans une grande population. Dans cet exemple, toute Nous venons d'illustrer l'application de la loi binomiale à un échantillon male réduite.

probabilités associées à l'intervalle approprié sous la courbe de la loi normale. Pour déterminer le niveau de confiance, on doit consulter une table des 95 %. Pour déterminer le niveau de confiance ne serait plus de devrait plus petit ou plus grand et le niveau de confiance ne serait plus de nous remplaçons le chiffre 1.96 par une autre valeur, cet intervalle devient qui contient la vraie valeur (inconnue) de  $p$  avec une probabilité de 95 %. Si l'estimateur habituel de  $p$ , pour ce genre de données, est  $\hat{p} = \bar{y} = \sum y_i / 5000$ . On définit  $s(\hat{p}) = \sqrt{\hat{p}(1-\hat{p}) / 5000}$ , où  $s(\hat{p})$  est notre estimation de l'erreur-typique de  $p$ . Or,  $\hat{p} \pm 1.96 s(\hat{p})$  est un intervalle aléatoire



## 2. MODÈLES À UNE VARIABLE

### 2.1 Modèles fondés sur la loi binomiale

Supposons que nous prélevons un échantillon dans une grande population et que nous observons une caractéristique de chaque unité choisie. Si la taille de l'échantillon est  $n$ , nous pouvons représenter les observations à l'aide de la notation  $y_1, y_2, \dots, y_n$ . Nous recueillons ces données parce que nous voulons faire des inférences sur la population à partir de cet échantillon. Par exemple, notre population pourrait être les personnes qui résident au Canada, et nos données peuvent être définies de la manière suivante:

$$y_j = \begin{cases} 1 & \text{si la personne } j \text{ est née au Canada} \\ 0 & \text{si la personne } j \text{ est née à l'extérieur du Canada.} \end{cases}$$

Nous voulons nous servir de cet échantillon pour déduire des informations sur la proportion de la population qui est née au Canada.

Si un échantillon aléatoire simple de  $n = 5,000$  résidents est prélevé et que la proportion réelle de personnes nées au Canada est  $p = 0.85$ , le nombre de personnes dans notre échantillon qui sont nées au Canada sera une variable aléatoire qui suit une loi binomiale exprimée par la fonction

$$f(y) = \binom{5000}{y} (0.85)^y (0.15)^{n-y} \quad y = 0, 1, \dots, 5000.$$

Comme nous savons que  $p = 0.85$ , nous pouvons fournir une description complète des propriétés de  $Y = \sum y_j$ , le nombre total de personnes nées au Canada dans notre échantillon. Par contre, dans la plupart des analyses statistiques, nous ne connaissons pas toutes les caractéristiques de la population mère et nous devons utiliser notre échantillon pour faire des inférences au sujet de cette population. Supposons que nous ignorons la valeur de  $p$  dans notre exemple. On peut alors prédire que le nombre de personnes nées au Canada dans notre échantillon sera une variable binomiale aléatoire dont la distribution correspond à la fonction

$$f(y) = \binom{5000}{y} p^y (1-p)^{5000-y} \quad y = 0, 1, \dots, 5000.$$

## INTRODUCTION AUX MODÈLES LINÉAIRES ET AUX MODÈLES LINÉAIRES GÉNÉRALISÉS

David A. Binder<sup>1</sup>

Cette étude décrit brièvement les modèles statistiques à une variable, les modèles de régression linéaire et les modèles linéaires généralisés. Des exemples d'une analyse de variance à deux facteurs et d'une régression logistique sont présentés.

### 1. INTRODUCTION

Notre propos ici est de donner un aperçu général de certaines notions utilisées en statistique pour construire des modèles applicables à des données.

L'estimation de moyennes et de proportions à partir de données sur un échantillon ordonné dans une population est devenue une pratique courante. À la section 2, nous abordons brièvement cette notion et nous décrivons les méthodes d'estimation d'intervalles de confiance.

Les modèles de régression linéaire et d'analyse de la variance sont souvent utilisés pour simplifier des données multidimensionnelles à l'aide d'un schéma comprenant un petit nombre de paramètres. Ces techniques sont des outils importants pour l'analyste qui veut approfondir sa compréhension d'un ensemble de données complexes. Nous examinons ces méthodes à la section 3.

Les notions utilisées dans les méthodes de régression linéaire peuvent être étendues à tout un éventail de problèmes, grâce aux modèles linéaires généralisés par Nelder et Wedderburn (1972). Ce type d'extension est particulièrement utile quand la variable dépendante est qualitative au lieu de continue. La section 4 porte sur la structure de ces modèles. Enfin, la section 5 offre une description très brève des diagnostics qui permettent de vérifier si un modèle est inadéquat et de déceler les anomalies linéaires.

- [8] Landis, J.R., Lepkowski, J., Eklund, S., et Stehouwer, S. (1982). A statistical methodology for analyzing data from a complex sample survey. Vital and Health Statistics, Series 2 - no. 92. DHHS Publ. no. 82-1366. Public Health Service, Washington, U.S. Government Printing Office.
- [9] Rao, J.N.K. (1984). Bootstrap inference with stratified samples. (non encore publié).
- [10] Rust, K.F. (1984). Techniques for Estimating Variances for Sample Surveys, The University of Michigan, thèse de Ph.D.
- [11] Verma, V., Scott, C., et O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. JRSS (a), 143, pp. 431-473.

## BIBLIOGRAPHIE

- qu'il y a trop de paramètres dans la double complexité des paramètres analytiques d'enquêtes complexes. Deuxièmement, les concepteurs de modèles ont tendance à faire disparaître cette complexité. Ils pourront toutefois nous aider à trouver des méthodes d'inférence plus exactes et compréhensives en améliorant l'utilisation et la présentation des paramètres analytiques d'enquêtes complexes.
- [1] Fay, R. (1982). Contingency table analysis for complex sample designs: Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 44-53.
  - [2] Kish, L. (1957). Some unsolved problems of complex samples. Communication présentée à une séance conjointe de l'American Statistical Association et de l'Institute for Mathematical Statistics.
  - [3] Kish, L., et Frankel, M.R. (1970). Balanced repeated replications for standard errors. JASA, 65, pp. 1071-1094.
  - [4] Kish, L., et Frankel, M.R. (1974). Inference from complex samples. JRSS (B), 36, pp. 1-74.
  - [5] Kish, L. (1980). Design and estimation for domains. The Statistician (Londres) 29, pp. 209-222.
  - [6] Kish, L., Groves, R.M., et Krotki (1976). Sampling errors for fertility. Occasional paper 17, Londres: World Fertility Surveys, 61 pages.
  - [7] Koch, G., Freeman, N., et Freeman, J. (1975). Strategies in the multivariate analysis of data from complex surveys. International Statistical Review, 43, pp. 53-59.

effets du plan pour les échantillons complexes (Kish et Frankel 1970 et 1974; Woodruff et Causey 1978). La méthode bootstrap pourrait également se répandre à l'avenir (Rao 1984).

Les paramètres analytiques comportent généralement des effets du plan supérieurs à 1 et ces résultats sont significatifs dans tous les sens de ce mot, mais ces effets du plan sont également moins élevés que ceux calculés pour les moyennes. Certaines régularités ont été mises en évidence dans les relations entre les effets du plan sur divers coefficients et les effets du plan sur les moyennes.

Pour nous aider à voir clair, il nous faut non seulement plus de travaux empiriques, mais aussi plus de résultats fondés sur la théorie de l'échantillonnage et l'analyse de modèles. Je dois avouer que je suis déçu que, depuis nos premiers travaux, nous n'ayons pas vu paraître plus de publications sur la théorie et les modèles qui seraient directement applicables à l'inférence à partir de données réelles. Il est nécessaire d'étudier les fondements empiriques des effets du plan, mais pour satisfaire notre besoin intellectuel de comprendre, il nous faut aussi une théorie mieux étayée et de meilleurs modèles. Même nos besoins pratiques ne sont pas complètement satisfaits par les travaux empiriques sur les effets du plan parce que les résultats dépendent à la fois des variables, du type d'estimations calculées pour le plan de sondage utilisé et de la population mère dans laquelle les données ont été recueillies. Cette source de variation à quatre dimensions est trop complexe et ce qui nous manque, c'est une théorie qui permet d'élaborer des modèles qui simplifient la réalité.

6. L'analyse de données qualitatives est un domaine important qui s'étend rapidement et plusieurs travaux ont été faits sur l'application de ces méthodes aux données d'enquêtes complexes (Fay 1982; Landis et coll. 1982; Koch et coll. 1975). Ces travaux abordent également des questions d'analyse de la variance qui étaient le point de départ de quelques-uns des premiers modèles, mais n'ont pas été approfondies par la suite (Kempthorne et Wilk 1955; Tukey et Cornfield).

7. Quant à l'avenir, j'ai bon espoir que de nouveaux éléments de théorie donneront naissance à des applications pratiques, sauf dans deux domaines. Premièrement, la statistique mathématique n'a pas et ne pourra pas nous donner une théorie complète des distributions qui soit directement utilisable parce



3. Les paramètres analytiques reposent sur la notion de sous-classes et de comparaison. Au cours des trois dernières décennies, beaucoup d'ouvrages et d'articles utiles ont été publiés à propos des variances et des effets du plan de sondage au niveau des sous-classes. On peut y trouver de nombreuses quantités de résultats empiriques, plusieurs observations pratiques basées sur ces résultats (Kish 1980, Kish et coll. 1976, Verma et coll. 1980) ainsi que des théories récentes (Rust 1984, chapitre 6).

a) On doit distinguer les domaines proprement dits des classes recoupées, qui sont plus fréquentes et constituent notre centre d'intérêt ici.

b) La probabilité de sélection demeure constante pour les classes recoupées, mais la taille des échantillons est très variable.

c) Les estimations de totaux et de moyennes calculées à partir d'échantillons complexes doivent être conservées sous la forme de quotients et d'expressions conditionnelles.

d) Les effets du plan au niveau des classes recoupées s'approchent d'une valeur presque égale à 1 proportionnellement à mesure que la taille des sous-classes par grappe primaire s'approche de 1. Ce modèle approximatif doit être utilisé avec prudence et sous certaines réserves, mais il est préférable à l'usage des échappatoires respectables qu'on trouve habituellement au sujet des effets du plan et selon lesquelles ces effets sont simplement soit égaux à 1 ou à une autre constante, soit identiques pour l'ensemble de l'échantillon. Un modèle adéquat peut souvent s'avérer meilleur qu'une série de calculs séparés et très variables.

4. Dans les comparaisons de moyennes appariées, les effets du plan sont souvent supérieurs à 1, mais beaucoup moins élevés que la somme de la variance de ces deux moyennes. Ce phénomène est attribuable à des covariances positives (et donc à une sorte d'additivité) et il a été observé régulièrement et dans des comparaisons de classes recoupées et des enquêtes périodiques (Kish 1965, 14.1: voir aussi les publications susmentionnées).

5. En ce qui concerne la production de paramètres analytiques, plusieurs méthodes exploitent les possibilités des ordinateurs électroniques. Les méthodes linéaires tayloriennes (méthodes delta), qui comprennent des programmes de différentiation automatisée, la méthode des duplications successives en blocs (balanced repeated replication) et la technique du jackknife se sont montrées régionales des outils importants pour l'estimation de la variance et des



erreurs d'échantillonnage. Au troisième rang, on retrouve les spécialistes en psychologie mathématique, en économétrie et en biométrie, qui, malheureusement, tirent leurs modèles linéaires directement de la statistique mathématique. Au quatrième rang, ce qui est encore plus malheureux, il y a les statisticiens mathématiques, qui ont tendance à oublier que "la fin ne justifie pas leurs moyennes," à invoquer l'hypothèse selon laquelle les observations sont indépendantes et distribuées identiquement ou à utiliser des formules d'exorcisme bayésiennes pour chasser les mauvais esprits hors du plan de sondage. Au cinquième rang, ce qui est le plus décourageant, on retrouve les théoriciens de l'échantillonnage qui construisent des théorèmes prouvant que, dans les modèles complètement spécifiés de superpopulations arbitraires, il n'est pas nécessaire de se soucier de la provenance et de la méthode de sélection des unités, ni de pondérer les observations pour tenir compte des inégalités des probabilités de sélection. Ces théoriciens ont même réussi à convaincre certains praticiens qu'ils peuvent rester sur l'Olympe avec leurs modèles sans jamais descendre sur terre. Là où se trouve la population.

Ces remarques forcément brèves vous permettent de constater que je suis extrêmiste pour plusieurs raisons: a) les effets du plan mesurés pour les paramètres analytiques révèlent généralement que les modèles des échantillons les mieux stratifiés sont mal définis; b) on observe souvent que les poids de sélection ont des effets sur les échantillons et c) il existe des relations entre les variables explicatives et les variables expliquées dans chaque individu, mais chaque individu fait partie de populations réelles et ces dernières ont des effets sur les plans de sondage (j'examine ces points en détail dans un ouvrage intitulé Statistical Design for Social Research, sur lequel je travaille actuellement et qui sera publié par Wiley en 1985.)

Ma philosophie est catégorique, mais, en pratique, je suis moins dogmatique. Je reconnais qu'en pratique: a) il n'est jamais possible de dénombrer complètement une population cible et, par conséquent, on doit toujours construire des modèles pour faire des inférences: b) l'échantillonnage probabiliste est trop coûteux et impraticable pour la plupart des expériences et c) malgré le manque de randomisation dans la sélection d'unités ou les analyses, on parvient souvent à des résultats fiables grâce à la prudence. À la répétition de l'échantillonnage, à la qualité du plan de sondage, à la propriété d'additivité et, aussi, un peu au jeu du hasard.

a. "Les paramètres d'un échantillon (moyennes, coefficients de régression etc.) approchent la valeur de la population mère à mesure que la taille de

l'échantillon augmente.

b. La convergence est généralement ralentie par les effets du plan.

c. Les effets du plan varient pour différents paramètres, pour différentes variables et pour différents plans de sondage" (Kish et Frankel 1974).

L'article cité plus haut présente également les arguments les plus convaincants à l'appui de ces propositions: la preuve en est d'ailleurs abondante (voir par exemple, Verma et coll. 1980). Néanmoins, deux statisticiens célèbres ont trahi complètement notre pensée dans des descriptions de notre article: "Les auteurs arrivent à la conclusion importante que les estimations d'intervalles de confiance d'un paramètre ne sont pas sensibles numériquement à la non-indépendance des observations introduite par les techniques d'enquête telles que l'échantillonnage stratifié en grappes". Hélas, cette erreur est citée par d'autres théoriciens qui n'ont pas lu notre réponse à l'intention des praticiens des méthodes d'enquête: "Ces auteurs n'ont pas du tout compris le message principal que nous ne cessons de répéter: les estimations d'intervalles de confiance sont très sensibles numériquement au manque d'indépendance entre les observations dans les techniques d'enquête complexes telles que l'échantillonnage stratifié en grappes" (Kish et Frankel, 1974).

Le fait que cette erreur soit partagée tant par des non-statisticiens que par des théoriciens de l'échantillonnage pose des problèmes à nous, praticiens des techniques d'enquête. Pour cette raison, nous travaillons à l'heure actuelle à fournir une explication plus claire de notre point de vue.

Faut-il tenir compte des erreurs d'échantillonnage dans le calcul de paramètres analytiques à partir des données d'enquêtes complexes? Ou est-ce que quelques-uns d'entre nous se sont consacrés à la solution d'un problème mathématique, voire insignifiant? Je comprends mieux saint Sébastien quand les flèches d'une foule de païens enragés lui pleuvaient sur le corps (même un acrobate se plaindrait). Au premier rang, il y a les spécialistes des études de marché et des maisons de sondage qui ne nous écoutent pas, bien que certains aient appris à mettre un  $\sqrt{\text{entre 2 et } (pq/n)}$ . Au deuxième rang certains économistes écrivent au à cause de leurs grands échantillons et de leurs grandes erreurs de mesure. Ils n'ont pas le temps de s'occuper des

théorie des distributions de statistiques analytiques doublement complexes n'a pas mobilisé les enthousiasmes des statisticiens mathématiques. Je comprends maintenant pourquoi il en est ainsi. La sagesse étant le fruit de l'expérience. Premièrement, les problèmes que les statisticiens, comme les autres scientifiques, tentent de résoudre sont choisis non en fonction d'un critère de nécessité, mais en fonction d'un critère de faisabilité à un moment défini. (La mise au point de bombes nucléaires est un bon exemple.) Deuxièmement, les problèmes de la théorie des distributions de paramètres d'échantillons complexes semblent trop difficiles à surmonter. Troisièmement, les solutions comporteraient elles-mêmes trop de paramètres pour être utiles. Le point de vue que j'ai exprimé en 1978 (Kish 1978) et qui est encore le même aujourd'hui est donc plus modéré. "Les nouvelles méthodes de calcul permettent d'obtenir des approximations des variances qui semblent satisfaisantes pour des besoins pratiques. Toutefois il serait préférable d'avoir une théorie mathématique des distributions de paramètres analytiques (tels que les coefficients de régression) qui est fondée non sur des hypothèses d'indépendance, mais sur des corrélations complexes entre les observations recueillies pour un échantillon. On peut oser espérer qu'un jour il y aura de nouvelles découvertes, mais les résultats ne seront pas universellement utilisables. à cause des complexités mathématiques, et surtout à cause du fait que le nombre de paramètres nécessaire sera trop grand pour les applications pratiques.

Je veux maintenant décrire sans ambages sept points importants au sujet des échantillons complexes. Ces points ne sont pas tous généralement connus ou acceptés, mais je vous demande d'y croire, de les mettre en pratique et de les enseigner, comme je le fais moi-même.

1. Les effets des plans de sondage complexes doivent être analysés séparément pour les estimations ponctuelles et les jugements de probabilité (comme, par exemple, les intervalles de confiance ou les tests d'hypothèses). Dans le cas des estimations ponctuelles, nous avons, pour tous les plans de sondage, une méthode cohérente pour évaluer les paramètres à l'aide des estimateurs pondérés selon la probabilité (estimateurs H-T). En revanche, les jugements de probabilité tels que les intervalles de confiance sont très sensibles aux effets du plan de sondage, surtout dans l'échantillonnage en grappes.

voir contraint à renoncer à une analyse qu'il considère indispensable. Par contre, s'il est trop impatient ou mal renseigné pour faire preuve d'une telle aléatoire simple qu'il peut trouver dans les manuels de statistique, ce qui conduit souvent à des erreurs très graves.

J'espère que les statisticiens mathématiques se rendront compte de l'importance des problèmes non encore résolus dans la recherche de paramètres analytiques pour les données recueillies à partir de plans de sondage complexes. Cette lacune cause plus d'erreurs fondamentales que tout autre écart des hypothèses habituelles.

Ces problèmes sont importants, ils n'ont pas encore été résolus et ils sont intéressants. Sommes-nous actuellement en mesure de proposer des solutions? Je crois que oui pour trois raisons. Premièrement, la théorie statistique a énormément progressé au cours des dernières années. Deuxièmement, l'augmentation rapide de la quantité d'ordinateurs électroniques et de leur qualité fait que le temps est propice pour la résolution de quelques-uns de ces problèmes. Troisièmement, on peut constater un nouvel intérêt à l'égard d'une méthode générale qui fait espérer des progrès rapides aboutissant à des approximations utiles. Au Survey Research Center, nous sommes en train d'élaborer cette méthode de calcul de la variance des coefficients de régression et d'autres paramètres pour lesquels des formules n'ont pas encore été mises au point.

Cette méthode me fait penser à la manière dont Alexandre a "résolu" le problème du noeud gordien. D'un point de vue théorique, je ne sais pas si cette méthode résout le problème ou le contourne. Mais, comme elle promet de bonnes approximations de variances qu'il est très important de mesurer, les statisticiens devraient l'accueillir avec enthousiasme et intérêt. Cette méthode permet de calculer des estimations d'intervalles de confiance de certains paramètres analytiques pour lesquels il n'existe pas de formule actuellement.

En ce qui précède est tiré intégralement d'un exposé que j'ai présenté au cours d'une séance conjointe de l'American Statistical Association et de l'Institute of Mathematical Statistics en 1957. La situation n'a guère évolué depuis la technique sur laquelle nos espoirs de trancher le noeud gordien ont été fondés en 1957 est souvent utilisée aujourd'hui. C'est la méthode des répétitions successives par blocs (balanced repeated replication) (Kish et Frane, 1970, 1974). Toutefois, mon discours passionné sur les lacunes de la



# SUR LE BESOIN DE PARAMÈTRES ANALYTIQUES POUR LES ÉCHANTILLONS COMPLEXES<sup>1</sup>

Leslie Kish<sup>2</sup>

Mon propos ici est de souligner l'importance et l'urgence de découvrir des expressions aléatoires utiles pour le calcul de paramètres analytiques adaptés aux plans de sondage complexes. Je voudrais m'adresser aux statisticiens mathématiques car ce problème devrait les intéresser puisqu'il en présente toutes les caractéristiques, c'est-à-dire qu'il est important, non encore résolu et soluble.

Il n'existe pas à l'heure actuelle d'analyse mathématique adéquate des problèmes les plus importants et les plus difficiles de la pratique de l'échantillonnage. Les manuels portent presque exclusivement sur le calcul de bonnes estimations d'agréats, de moyennes de variables et de moyennes de quotients. Il est parfois question de la différence entre deux valeurs d'un de ces paramètres, mais ce problème n'est abordé que de façon passagère et sporadique. C'est à cela que se limitent nos outils statistiques pour l'étude d'échantillons complexes.

Depuis la naissance de la théorie de l'échantillonnage, les méthodes probabilistes ont acquis progressivement un rôle privilégié dans le cours des travaux d'échantillonnage, et ces méthodes sont appliquées à des plans de sondage qui sont à la fois économiques et complexes. Le volume de données de haute qualité recueillies au moyen d'enquêtes n'a cessé d'augmenter, grâce à ces techniques, et les spécialistes veulent soumettre leurs données empiriques à des analyses de plus en plus compliquées. Toutefois on ne dispose pas de la théorie statistique mathématique requise pour faire des jugements valables. Pour utiliser les paramètres analytiques habituels, on doit supposer que les éléments choisis sont indépendants, mais cette indépendance est absente dans les plans de sondage complexes. Ainsi, un chercheur peut se

<sup>1</sup> Exposé présenté au symposium.  
<sup>2</sup> Leslie Kish, Institute for Social Research, The University of Michigan





## PRÉFACE

Ce numéro renferme des communications présentées lors d'un colloque intitulé **l'Analyse de données d'enquête - aspects et méthodes** qui a eu lieu à Statistique Canada le jeudi 3 mai 1984.

Ce colloque a été parrainé par le Comité d'étude des méthodes de Statistique Canada et le Laboratoire de recherche en statistique et probabilité des universités Carleton et Ottawa. Il avait pour objet de démontrer comment des réalisations récentes en matière d'analyse de données provenant d'enquêtes complexes pourraient être appliquées aux études analytiques effectuées à Statistique Canada.

Il a débuté par des observations du Statisticien en chef, Martin B. Wilk, qui a souligné l'importance que Statistique Canada attache à l'amélioration de ses moyens de recherche et de développement et aux efforts communs des praticiens et des universitaires à ce sujet. Le colloque comprenait deux séances: une le matin, présidée par Leslie Kish de l'Institut pour la recherche sociale de l'Université de Michigan, au cours de laquelle des communications de Statistique Canada ont été présentées par D. Binder, P. Cholete, L. Heslop, et S. Kumar, en plus d'un aperçu concernant l'analyse présentée par le président.

La séance de l'après-midi, présidée par le sous-statisticien en chef, Ivan Fallegi, s'est ouverte par de brèves observations du président, et des communications présentées par R. Fay, U.S. Bureau of the Census et W. Fuller, de l'Université de l'Iowa. La séance s'est terminée par une discussion générale des faits nouveaux concernant l'analyse de données, sous la direction de J.N.K. Rao, de l'Université Carleton. Plus de 200 représentants d'universités et de ministères et organismes fédéraux et provinciaux ont participé au colloque.

On trouvera également à la fin une bibliographie sélective sur le sujet établie par l'équipe de projet de l'analyse des données d'enquêtes complexes.



## Édition spéciale

Une revue préparée par Statistique Canada

## Comité de rédaction:

R. Platek  
M.P. Singh  
G.J.C. Hoie  
C. Patrick  
P.F. Timmons  
H. Lee  
- Rédacteur en chef  
- Président  
- Rédacteur adjoint

## Politique de la rédaction

La revue "Techniques d'enquête" veut donner aux personnes qui s'intéressent aux aspects pratiques de la conduite d'enquêtes, la possibilité de publier sur ce sujet dans un cadre canadien. Les textes pourront porter sur toutes les phases de l'élaboration de méthodes d'enquêtes: les problèmes de conception causés par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada.

## Présentation de documents pour publication:

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, 4e étage, Édifice Jean Talon, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi.



Édition spéciale

Une revue préparée par Statistique Canada

TABLE DES MATIÈRES

Préface.....	0
Sur le besoin de paramètres analytiques pour les échantillons complexes.....	1
LESLIE KISH.....	1
Introduction aux modèles linéaires et aux modèles linéaires généralisés.....	10
DAVID A. BINDER.....	10
L'ajustement des séries infra-annuelles aux régressions annuelles.....	39
PIERRE A. CHOLETTE.....	39
Analyse des dépenses consacrées à l'énergie.....	54
LOUISE A. HESLOP.....	54
Régression logistique et analyse de données de l'enquête sur la population active.....	68
S. KUMAR and J.N.K. RAO.....	68
Application de modèles linéaires et log-linéaires aux données d'échantillons complexes.....	90
ROBERT E. FAY.....	90
Application de la méthode des moindres carrés et de techniques connexes aux plans de sondage complexes.....	107
WAYNE A. FULLER.....	107
Bibliographie sélective pour l'analyse des données d'enquêtes complexes.....	131







# TECHNIQUES D'ENQUÊTE

juin 1984

volume 10

numéro 1

ÉDITION SPÉCIALE

analyse des données d'enquête  
— problèmes et méthodes

Une revue préparée  
par Statistique Canada

Canada

12-001



Statistics Canada Statistique Canada

Collection  
Publications

# SURVEY METHODOLOGY

A JOURNAL  
OF  
STATISTICS CANADA



VOLUME 10, NUMBER 2  
DECEMBER 1984

Canada



## Order Coupon

Mail to: Publication Sales and Services  
Statistics Canada  
Ottawa, Ontario  
Canada K1A 0T6

Please enter my subscription to the Survey Methodology Journal (Cat. no. 12-001). The price is \$10.00 per copy, \$20.00 per year in Canada, \$11.50 per copy, \$23.00 per year for other countries (payment to be made in Canadian funds or equivalent).

I am currently a subscriber to Statistics Canada Publications:

☐ Yes. My subscriber reference number is \_\_\_\_\_  
(Please await an invoice before paying.)

☐ No. Your order cannot be processed unless a purchase order is enclosed or the method of payment is indicated by checking one of boxes below.

Purchase Order No. \_\_\_\_\_ (Please enclose purchase order)

Number of Copies \_\_\_\_\_ Amount Sent \$ \_\_\_\_\_

☐ My remittance payable to the Receiver General for Canada/Publications is enclosed.  
(INTRA No. 0540) (Creditor Account No. 0051)

☐ Charge to my Statistics Canada Account, No. \_\_\_\_\_

☐ Charge to my ☐ VISA ☐ MASTERCARD.

Card No. \_\_\_\_\_

Expiration Date \_\_\_\_\_

Name of Card Holder \_\_\_\_\_  
(Please print)

Issuing Bank \_\_\_\_\_

Signature of Card Holder \_\_\_\_\_

Ship to (please print):

Name (Organization) \_\_\_\_\_

Department \_\_\_\_\_

Attention \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_

Province \_\_\_\_\_

Postal Code \_\_\_\_\_

Telephone \_\_\_\_\_



Statistics Canada

# SURVEY METHODOLOGY

JOURNAL OF STATISTICS CANADA

December 1984

Published under the authority of  
the Minister of Supply and  
Services Canada

Minister of Supply  
and Services Canada 1985

April 1985  
0-200-501

Price: Canada, \$10.00, \$20.00 a year  
Other Countries, \$11.50, \$23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 10, No. 2

SN 0714-0045

Ottawa





# SURVEY METHODOLOGY

A Journal of Statistics Canada  
Volume 10, Number 2, December 1984

## CONTENTS

M.P. SINGH, J.D. DREW, and G.H. CHOUDHRY Post '81 Censal Redesign of the Canadian Labour Force Survey.....	127
D.A. BINDER, M. GRATTON, M.A. HIDIROGLOU, S. KUMAR, and J.N.K. RAO Analysis of Categorical Data from Surveys with Complex Designs: Some Canadian Experiences.....	141
P.A. CHOLETTE Estimating Economic Cycles in Semi-Annual Series.....	157
I. COULTER The Use of Matching in the Evaluation of Non-Sampling Errors in the 1981 Canadian Census of Agriculture .....	165
A. CHAUDHURI and R. MUKHERJEE Unbiased Estimation of Domain Parameters in Sampling without Replacement .....	181
D. DOLSON, P. GILES, and J.-P. MORIN A Methodology for Surveying Disabled Persons Using a Supplement to the Labour Force Survey .....	187
Corrigendum .....	199
Acknowledgements .....	201

# SURVEY METHODOLOGY

A Journal of Statistics Canada

## EDITORIAL BOARD

<b>Chairman</b>	R. Platek	Statistics Canada
<b>Editor</b>	M.P. Singh	Statistics Canada
<b>Associate Editors</b>	K.G. Basavarajappa	Statistics Canada
	D.R. Bellhouse	University of Western Ontario
	E.B. Dagum	Statistics Canada
	J.F. Gentleman	Statistics Canada
	G.J.C. Hole	Statistics Canada
	T.M. Jeays	Statistics Canada
	G. Kalton	University of Michigan
	C. Patrick	Statistics Canada
	J.N.K. Rao	Carleton University
<b>Assistant Editor</b>	C.E. Särndal	University of Montreal
	V. Tremblay	University of Montreal
<b>Assistant Editor</b>	H. Lee	Statistics Canada

## MANAGEMENT BOARD

R. Platek (Chairman), E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

---

## EDITORIAL POLICY

The Survey Methodology Journal will publish articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, smoothing and extrapolation methods, demographic studies, data integration and analysis and related computer system development and applications. The emphasis will be on the development and evaluation of specific methodologies as applied to actual data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

## Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit the manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

**To Mr. Paul Francis Timmons**

in recognition of his contributions as a founding member  
of the Editorial Board until his retirement.



## Post '81 Censal Redesign of the Canadian<sup>1</sup> Labour Force Survey

M.P. SINGH, J.D. DREW, and G.H. CHOUDHRY<sup>1</sup>

### ABSTRACT

Following each decennial population census, the Canadian Labour Force Survey (CLFS) has undergone sample redesign to reflect changes in population characteristics and to respond to changes in information needs. The current redesign program which culminated with introduction of a new sample at the beginning of 1985 included extensive research into improved sample design, data collection and estimation methodologies, highlights of which are described.

KEY WORDS: Continuous survey; Multi-stage sample design; Stratification; Sample reallocation; Telephone interviewing; Raking ratio estimation.

### 1. INTRODUCTION

The Canadian Labour Force Survey (LFS), the largest monthly household survey conducted by Statistics Canada, has been redesigned in the past following each decennial census. As a part of 1981 post censal redesign, an intensive program of research as outlined in an earlier paper (Singh and Drew 1981a) was undertaken in the areas of sampling, estimation and data collection methodologies. As the reliability of labour market data at the national and provincial levels was of sufficiently high standard, the major emphases in this redesign were on improving the reliability of subprovincial data and on making the survey more cost efficient. Towards the latter, the main thrusts were on increased automation of various steps involved in sampling, greater use of Census data in place of independently obtained information in updating the sample, and increased telephone interviewing as a regular feature of the survey. As for the improvement in the subprovincial data, alternative sampling and estimation methods were investigated leading to refinements in the earlier methods, coupled with reallocation of the sample within provinces.

This paper presents an overview of the findings of various theoretical and empirical investigations and field tests undertaken during the redesign program. Sections 2 and 3 provide the background information on objectives and the old design, while Sections 4, 5 and 6 highlight the findings of investigations leading to changes in sampling and data collection methodologies. Section 7 deals with estimation issues, and sample reallocations are discussed in Section 8. Implications of the changes made in the redesigned sample on other associated surveys are outlined in Section 9, and finally in Section 10 the benefits from the major improvements in the redesigned sample are briefly recounted, along with some mention of future research plans.

### 2. OBJECTIVE SETTING

A fundamental step in the redesign of a recurring survey is the re-establishment of survey objectives. For the LFS, this involved reassessment not only of the survey's principal role as provider of current labour market information, but also of its use as a central vehicle within Statistics Canada for conducting household surveys (Singh and Drew 1981b).

<sup>1</sup>Presented at ASA meetings, Section on Survey Research Methods, Philadelphia, August 1984.

<sup>1</sup>M.P. Singh, J.D. Drew, and G.H. Choudhry, Census and Household Survey Methods Division, Methodology Branch, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.



At the early stages of the redesign program it was decided that while primary orientation towards the LFS should be maintained, efforts should also be made to enhance flexibility of the vehicle for general applications. In this light, several changes are being adopted that will benefit not only the LFS but its associated surveys as well. Requirements for the labour market data are noted below, while highlights of the changes resulting in the improvements for associated surveys are given in Section 9.

Objectives relating to provision of labour market data were established in consultation with the statistical focal points within each of Canada's ten provinces, and with key federal and provincial departments, through annual Federal/Provincial Conferences on Labour Statistics and bilateral follow-ups. In general, these consultations revealed satisfaction with data reliability for provincial and national data, but a strong desire for improved subprovincial data. Specific data reliability objectives adopted for the redesigned sample are as follows:

- (i) for Canada and each of the ten provinces, no reduction in current reliability for monthly estimates of level and estimates of month-to-month change in total employment and unemployment.
- (ii) for the 24 Census Metropolitan Areas as defined by the 1981 Census, monthly estimates of unemployed with coefficients of variations (CV's) of 20% or less.
- (iii) for 66 subprovincial Economic Regions agreed to in consultation with the provinces, monthly estimates of unemployed with CV's of 25% or less.
- (iv) for cities with population of 60,000 or more in Quebec and Ontario, and 25,000 or more in the remaining provinces, quarterly estimates of unemployed with CV's of 25% or less.

Attainment of these objectives necessitated a reallocation of sample within the provinces that entailed a shift of sample from larger cities and Economic Regions to smaller ones. This, coupled with the desire to reduce the cost of the LFS as discussed in Section 8, created high expectations from the research projects for identifying more cost efficient strategies for data collection and production of statistics from the LFS. In the following sections, these issues are addressed for the two main parts of the survey namely, Self-Representing (SR) and Non Self-Representing (NSR) Areas.

### 3. OLD LFS DESIGN

A complete description of the old LFS design is given by Platek and Singh (1976). Salient features are noted in this section to provide a context for discussions in later sections.

The Self-Representing Units (SRUs) in the old design corresponded to those cities which, when designing the survey, were sufficiently large to yield an expected sample of 20 dwellings, the minimum thought acceptable for an interviewer. Minimum SRU sizes ranged from a population of 10,000 in the Atlantic Region to 25,000 in Quebec and Ontario.

Within large SRUs, deep geographic stratification was carried out by grouping together contiguous Census Tracts — geostatistical areas with populations in the 3,000 – 5,000 range whose stability from one Census to the next makes them a convenient operational unit — without any regard to optimality of characteristics. Primary sampling units, called clusters, corresponding approximately to city blocks, were delineated on the basis of field counts obtained in 1971. A two stage sample of clusters and dwellings was selected following the Rao, Hartley and Cochran (1962) pps random group method. In addition to the area frame, an open-ended list frame of apartments was maintained in the larger cities.

A major advantage of the selection method for the area frame lies in its flexibility for changing in the sample size (Singh and Drew 1977), and for sample updating (Platek and Singh 1977, Drew, Choudhry, and Gray 1978). Sample updating is necessary in SRUs because over time the design counts used in the pps selection become out of date, leading to higher sampling variance for survey estimates. Since sampling is done independently in each random group, a Keyfitz (1957) sample update can be carried out, under which revised probabilities of selection based

on recent dwelling counts can be incorporated, while maximizing retention of already selected units and avoiding overlap of selected dwellings between the pre- and post-update-samples. Regular updating of high growth SR areas occurred from 1978 until the beginning of the redesign period in 1982, during which time almost half of the frame was updated. While the intensity of updating was not sufficient to reduce survey design effects to levels experienced during the first 4 years of the survey, it was sufficient to arrest further deterioration which had been averaging 7 - 8% per year for unemployed.

The Non-Self Representing (NSR) units are the areas outside the SRUs comprised of rural areas and smaller urban centers. In NSR areas, stratification based on industry classifications was carried out within Economic Regions, subject to the restriction that strata should be contiguous land areas. Within strata, Primary Sampling Units (PSUs) were delineated such that each PSU represented its stratum to the extent possible with respect to the ratio of rural to urban population and important LF characteristics. While the rural portions of a PSU were comprised of collections of contiguous rural EAs, urban portions could not always be made contiguous to the rural due to the restrictions placed on maintaining the rural to urban population ratio. Frequently larger urban centres had to be shared amongst several PSUs within the stratum.

At the time of the 1973 Redesign, two PSUs were selected per stratum using the randomized pps systematic method (Hartley and Rao 1962). In 1977, the LFS sample size was increased from 33,000 to 55,000 households per month, with the additional sample being allocated so as to improve data reliability at the province level. Thus the smaller provinces received larger proportionate sample size increases. The increase was achieved in NSR areas by selecting 1 - 4 additional PSUs per stratum (Gray 1975).

Within selected PSUs, urban and rural portions were sampled independently. In the urban portion of the selected PSU a two stage sample of clusters and dwellings was selected, whereas in the rural portion of the PSU a three stage design was used with secondaries (which are Census EAs or combinations), clusters (identifiable land areas having up to 20 dwellings) and dwellings as the sampling units. Except for the last stage, randomized pps systematic sampling was used in the selection.

#### 4. REDESIGN OF THE SELF REPRESENTING AREAS

The criterion for cities to qualify as SRUs in the new design was raised to a minimum sample of 50 dwellings, since analysis of cost data indicated significantly higher unit costs for smaller SR assignments. The composition of the SR universe remained largely unaffected, however, due to the off-setting influences of the increase in the LFS sample size from 33,000 to 55,000 households during the late 1970's, and due to the reallocation of the redesigned sample.

For reasons noted earlier, the basic design in the SR areas remained the same and the main thrust for these areas was to update the size measures without resorting to a costly independent field count as used in the last redesign. In order to exploit the data collected during the 1981 census for this purpose different approaches were used in the block-faced cities (larger cities where data were available at block face level) and non-block-faced cities. The choices of updating method and sampling units by the two approaches are given below, whereas the two level stratification adopted for SR areas is given in Section 5.

##### **Block-Faced Cities**

The availability of Census data at the block face level in the built-up portions of the larger cities was the key factor in deciding to completely redesign the sample in these cities, which account for  $\frac{2}{3}$  of the SR frame. The redesign of these cities also provided the opportunity to introduce improved stratification as described in the next section.

For block-face portions of the cities, Census blocks were adopted as clusters (i.e., PSUs). Variance components under a two stage RHC random group design were studied for different choices of clusters — Census Blocks or EAs — by simulating the LFS design using 1976 Census data for the SRUs of Halifax and Saskatoon (Choudhry, Drew, Lee 1984). Study results, for the case of up-to-date size measures, showed little difference between the EA and block in terms of sampling variances. Hence, the decision in favour of blocks was made on the basis of operational considerations. The blocks provided a ready made frame (with splitting or combining needed in only 5 – 10% of cases), permitting a highly automated design with very low redesign costs. Importantly as well, data for future Censuses will be retrievable for the geostatistically stable blocks (but not for EAs which as operational units change at each Census), permitting low cost quinquennial updating of the sample. The built-up portions comprise 86% of these cities.

The EA was adopted in the non-block faced portions of the cities. Whereas the study results considered only the up-to-date case, it was felt that the EA being a larger unit than the block would be more robust to the highly clustered growth which can occur in the non-built-up portions of cities. Also, adoption of the EA in these areas permitted very low redesign costs, since roughly 80% of these areas fell in Quebec and Ontario where due to the low sampling rates very little splitting was needed.

The variance study results, combined with cost results from the Time and Cost Study (Lemaître 1983), showed variances per unit cost to be quite flat in the range of 2 – 8 selected dwellings per cluster. Hence it was decided to retain the density of 4 – 5 dwellings per cluster used in the old design in strata where clusters were blocks, but to increase the density to 6 – 8 dwellings in the case of the EAs, due to their larger sizes.

### **Non-Block-Faced Cities**

Since over 70% of the non-block-faced SRUs were either new or had changed boundaries, and since most of those remaining had not been updated since the 1973 redesign, it was decided to completely redesign these cities. Clusters were taken as individual or combined Census blocks in the built-up portions of the cities, with the dwelling counts obtained directly from visitation records and maps completed by Census enumerators. In the non-built-up portions, EAs or split EAs were taken as clusters, with field counts sometimes being required to do the splitting.

The use of Census visitation records, while more costly than procedures followed in the block-faced areas, nevertheless resulted in significant cost savings over the procedure followed in the old design of obtaining independent field counts.

## **5. STRATIFICATION**

### **5.1 Algorithm and Choice of Stratification Variables**

A modified version of a non-hierarchical algorithm due to Friedman and Rubin (1967) was adopted for stratification purposes in both SR and NSR areas, on the strength of findings reported by Judkins and Singh (1981), and Kostanich, Judkins, Singh and Schautz (1981) who evaluated several stratification algorithms for use in the U.S. Bureau of the Census' Current Population Survey. New features incorporated into the algorithm were a capacity to form geographically contiguous and/or compact strata, and the option to form either homogeneous clusters (i.e., strata) or heterogeneous clusters (i.e., primary sampling units within NSR strata). A detailed description of the method, and results of empirical evaluation studies are given by Foy, Bélanger, Drew and Joncas (1984). An overview is presented below.

The algorithm starts with a random partitioning of units into a specified number of strata. An iteration consists of examining in turn each stratification unit and moving it to any stratum which will reduce a weighted multivariate within stratum sum of squares, while continuing to satisfy constraints on strata sizes. The algorithm converges at a local optimum when moving any one of the units would increase the within strata sum of squares. Based on the findings of Judkins and Singh (1981), local optima were improved upon by the use of a moderately large number of random starts (i.e., 30).



For the contiguity option, a matrix is inputted specifying for each unit, all others contiguous to it. Contiguous initial strata respecting the size constraints are then built up from units chosen as random starting points. During the optimization step, an extra condition is imposed on transfer of units that contiguity be maintained. To achieve compactness, population centroids (longitude and latitude) are added as variables in the weighted sum of squares to be minimized.

For both NSR and SR areas, a multivariate stratification has been carried out using 1981 Census data for up to 17 stratification variables. Population variables include: total employed; employment income; persons with secondary education; population 15+, 15-24 and 55+; and labour force in agriculture, forestry-fishing, mining, manufacturing, construction, transport, and services. Dwelling related variables include: total dwellings, dwellings rented, one person households, and two persons households.

Any industry variables accounting for less than 2% of the labour force of the area being stratified were dropped and other variables were scaled to be equally important in the optimization process. Unemployed was not included as a stratification variable due to its instability. Study findings showed that strata formed without unemployed when evaluated at the next Census were more efficient not only for other characteristics, but for unemployed as well. The inclusion of the neighborhood type variables, on the other hand, did result in improved efficiency for unemployed.

## 5.2 Two Level Stratification in SR Areas

In the larger SRUs with sample sizes of 300 or more households, two levels of stratification were adopted. Primary strata, with sample yields of 150 - 170 households from the area and apartment samples combined, are comprised of collections of geographically contiguous Census Tracts. As such, primary strata are designed to correspond to two interviewer assignments. Three to four non-geographic secondary areal strata each yielding six or a multiple of six sampled clusters are formed within primary strata, with Census Tracts as stratification units, and with optimization based on the 1981 Census characteristics of non-apartment dwellers.

Apartments are sampled separately, in the form of a pps systematic sample from an open-ended list, which generally comprises a single stratum for the entire SRU. Sorting of the apartments existing at the time of design by primary strata yielded an implicit geographic stratification to the apartment sample.

In the smaller block-faced SRUs which warranted neither separate apartment sample nor geographic primary strata, optimal non-geographic areal strata were formed directly. In the non-block-faced cities, with considerably less scope for stratification, simple geographic strata were opted for.

The two levels of stratification in the larger SRUs had appeal on both operational and technical grounds. The relaxing of geographic constraints over those existing in the old design permitted greater optimality to be achieved, while the retention of contiguity at a higher level will provide a suitable unit for sample updating purposes later in the decade, and will facilitate the planning of interviewer assignments. Also, in the old design, SR strata were likely to be covered entirely by a single interviewer, and hence the variance estimates did not reflect the correlated response variance component of total variance. To the extent that within strata, interviewer assignments are geographic and secondary strata are non-geographic, an interpenetration of strata and interviewer assignments will be achieved in the new design without incurring any additional data collection costs, resulting in this component being better reflected in the variance estimates.

Table 1 presents study results for two SRUs — Ottawa and Quebec City — comparing efficiencies of the geographic strata used in the old design with those of optimal two-level strata, formed using 1971 Census data. Percent reductions in the first stage variance due to stratification, calculated at the time of the 1981 Census, indicate largest improvements under the optimal stratification for income and rented dwellings. The only marginal gains for other characteristics including employed and unemployed point to the strength and robustness of the simple, but deep, geographic stratification in the old design.

**Table 1**  
 % Reduction in First Stage Variance Due to Stratification  
 Old vs New Methods

Variable	Stratification Method		Variable	Stratification Method	
	old	new		old	new
total employed	9.1	12.6	agriculture <sup>1</sup>	5.9	3.9
employment income	18.1	30.4	forestry/fishing <sup>1</sup>	3.1	2.4
secondary education	39.4	42.1	mining <sup>1</sup>	4.8	3.0
population 15+	9.2	12.6	manufacturing	23.5	23.1
population 15 - 24	12.9	17.6	construction	11.9	11.2
population 55+	25.3	29.7	transport	4.2	6.4
total dwellings	28.5	33.1	services	14.5	19.8
dwellings rented	20.9	28.8	unemployed <sup>1</sup>	7.1	9.7
1 person households	33.7	38.4			
2 person households	27.5	29.6			

<sup>1</sup> characteristics not used in optimization for new method

In the NSR areas, the same clustering algorithm was used within each Economic Region to form either rural, or mixed urban and rural strata, depending on the design adopted, as discussed in Section 6.3. Also the adaption of the clustering algorithm for use in PSU formation is described in Section 6.5.

## 6. DESIGN CONSIDERATIONS IN NSR AREAS

### 6.1 Extension of Telephone Interviewing

Telephone interviewing for months 2-6 in the sample was introduced during the early 1970s in Self Representing Areas, primarily to reduce cost. However in NSR areas, all interviewing continued to be done in person due to concern over the high instance of party lines vis-à-vis the confidentiality of the data being collected. Nevertheless, in recognition that not only immediate cost benefits from telephoning were at issue, but so also were the longer term potential benefits from the use of new technologies such as Random Digit Dialing and Computer Assisted Telephone Interviewing (CATI), it was decided to test the feasibility of extending telephone interviewing to NSR areas.

A first field test was restricted to urban areas having over 80% private lines. The test was conducted on a portion of the actual LFS sample, with the principal objective of assessing the data quality implications of telephone interviewing. To facilitate this analysis, interview assignments were split between the telephone and personal procedure.

This test ran from January 1982 to June 1983 with a gradual phase-in to ensure no adverse impact on the ongoing survey. Principal findings were: lower non-response rates for the telephone sample (3.4% versus 4.3% for the control sample); a high instance of households with telephone (96% for all provinces but one); a low instance (1%) of households not agreeing to telephone interviewing; and no detectable differences in estimates for labour force characteristics.

A second test carried out in the rural areas had comparable findings. Based on the positive findings from both tests, the decision was taken to introduce telephone interviewing across the board in NSR areas during the remainder of 1983 and early 1984.

The decision to extend telephone interviewing had the following principal implications on the design of the NSR sample:

- (i) Increase in assignment sizes: In the old design, NSR assignment sizes averaged 50 dwellings. Evidence that per unit costs were lower for larger assignments (Lemaître 1984), and the reduction in travelling under telephoning, both supported increasing the design yield per NSR PSU to 55 – 60 dwellings.
- (ii) Level of assignment of rotation numbers: Unlike the old design, in the new design, all dwellings within secondaries will receive the same rotation number, which will cut down on the number of visits to the secondaries in month 2 – 6 in the sample.

## 5.2 Elimination of Stage of Sampling in Rural Areas

In the old design, the rural sample within PSUs was selected in three stages: secondaries (Census Enumeration Areas), clusters, and dwellings. The clusters corresponded to identifiable land areas containing up to 20 dwellings, which were delineated on the basis of field counts obtained whenever a new secondary entered the sample. Within secondaries, generally 5 – 6 clusters, with 3 – 4 dwellings per cluster, were selected.

The rural cluster stage was identified early in the redesign program as a possible candidate for elimination, on the grounds that (i) the sample variance would be reduced due to having one less stage of samplings, and (ii) the lead time required to introduce new secondaries into the sample could be shortened from 13 months to 7 months.

A field study was carried out on a sample of secondaries entering the IFS sample, in order to assess the feasibility of maintaining good quality dwelling lists for entire rural EAs, and to examine costs under such a procedure, with positive results on both counts. The variance implications of eliminating the cluster stage were also studied. Using 1971 Census data to simulate both the old and alternative design, components of variance were obtained for the Horwitz-Thompson estimator without ratio estimation. The percent reduction in total variance under the alternative design was found to range from 20 – 25% for major labour force characteristics (Choudhry, Lee, and Drew 1984).

On the basis of these findings, an early decision was taken to eliminate the rural cluster stage of sampling, and attention was turned to more global aspects of the NSR design.

## 5.3 Design with Urban/Rural Stratification

The old design featured implicit urban/rural stratification. PSUs were formed to have approximately the same ratio of urban to rural population as the stratum, and within selected PSUs the urban and rural portions were sampled independently. A premise underlying the design was that the PSU should correspond to an interviewer's assignment. However, in practice this correspondence was weakened since in order to attain the desired urban – rural ratio, frequently the urban and rural portions of PSUs were not contiguous.

As an alternative to the old design,  $D_0$ , (with the rural cluster stage eliminated), a design,  $D_1$ , featuring explicit urban/rural stratification was studied. Like  $D_0$ , the alternative design  $D_1$  consisted of 3 stages of sampling in both urban and rural areas. In urban strata, the stages were: PSUs (consisting of individual or nearby urban centers), clusters, and dwellings. In rural strata, the stages were: PSUs, (consisting of collections of nearby rural EAs), secondaries (EAs), and dwellings. Under  $D_1$ , both urban PSUs and rural PSUs were designed independently to yield samples corresponding to interviewer assignments.

The two design alternatives were evaluated, from the point of view of variance and cost (Choudhry, Drew, Lee 1984). In the variance study both designs were simulated for the case of PSUs per stratum using design counts based on the 1971 Census, and study variables based on 1976 Census data.

In terms of costs, a simple model was developed for  $D_0$ , the old design under telephone interviewing, and components were estimated using results from a detailed Time and Cost Study (Lemaître 1983). Relative costs for the travel components between designs  $D_0$  and  $D_1$  were estimated by means of a simulation study, in which average dispersion of the sample under the two designs was obtained up to the second stage of sampling using the population centroids of the EAs.



Findings were that the design  $D_1$  was 1.09 times as cost efficient as  $D_0$ , and that from the combined perspective of cost and variance,  $D_1$  outperformed  $D_0$  with overall efficiencies of 1.25 for employed and 1.05 for unemployed.

Based on these findings, design  $D_1$  was adopted in 70% of Economic Regions with sufficient urban and rural population to yield separate strata. In the remaining Economic Regions, with the exception of Prince Edward Island, design  $D_0$  was adopted (see Section 6.6).

#### 6.4 Number of PSUs Selected Per Stratum

In the LFS design, since the sample yield per PSU is fixed, the number of PSUs selected per stratum also determines the number of strata. In over two thirds of cases, the urban, rural or combined strata within ERs yielded only enough sample for 2 or 3 PSUs. Further stratification in these cases was ruled out on the grounds that there should be at least 2 PSUs per stratum to permit unbiased estimation of variance.

The remaining ERs were stratified to the point of 2 or 3 PSU's per stratum. Estimated first stage variance reductions over the situation under the old design of from 3 – 6 selected PSU's per stratum were up to 14% for employed (Choudhry, Lee, and Drew 1984). The stratification was carried out using the clustering algorithm described in Section 5.

#### 6.5 Use of Clustering Algorithm in Formation of NSR PSUs

In both the old and new LFS, stratification is carried out prior to formation of NSR Primary Sampling Units. PSUs are delineated within the stratum to be as similar as possible with respect to stratification variables, while being as geographically compact as possible. PSU delineation which was carried out using the clustering algorithm noted earlier, required minimization of geographic and maximization of the non-geographic variables.

#### 6.6 Two Stage Design for Prince Edward Island

For Canada's smallest province, Prince Edward Island, sampling rates are 4% in order to produce monthly LF estimates with required levels of data reliability. In view of the high sampling rates, a less clustered design consisting of a two stage sample of EAs and dwellings, with deep geographic stratification was adopted. It was found to have marginally higher costs than  $D_0$ , however from the overall perspective of cost and variance, it came out well ahead with efficiencies of 2.21 and 1.11 for employed and unemployed relative to  $D_0$  (Choudhry, Lee, and Drew 1984).

### 7. ESTIMATION

#### 7.1 Final Stage Ratio Estimation

In the old LFS, a final stage ratio estimation was carried out by detailed province/age/sex cells. With the development within Statistics Canada of improved and more timely subprovincial population estimates (Verma, Basavarajappa, and Bender 1982), an intermediate ratio estimation step was studied in which survey estimates of population 15+ for subprovincial areas are ratio adjusted to external estimates prior to the usual final ratio estimation. Findings were that the procedure, while not impacting on the variances of provincial level data, resulted in variance reductions for sub-provincial areas ranging from close to 70% for employed to 7% for unemployed (Earwaker and Bélanger 1981). In practice a raking ratio procedure in which the two ratio estimation steps are iterated until both marginal controls are satisfied was adopted, beginning in 1983.

#### 7.2 Improved Estimates for Household and Family Units

Paul and Lawes (1982) used LFS longitudinal data files, which link households over the six months in the sample, to demonstrate that non-response rates are higher amongst households with fewer members. For the old LFS, non-response adjustment consisted of re-weighting at local area levels. This was done without regard to household size, hence the resulting estimates of households and families by size had 1 – 3% biases. Another problem related to the inconsistency

of family and individual based statistics (Macredie 1983). When demographic estimates of families by size, currently under development by Statistics Canada's Demography Division, become available, it is intended to incorporate them as an extra dimension in the final stage raking ratio estimation procedure, to address both problems.

As an interim measure, the use of LFS longitudinal data is being studied as a mean to derive household size distributions based on both respondents and non-respondents, prior to the final raking ratio estimation (Ghangurde 1984).

### 7.3 Small Area Estimation

Demand for Labour Force estimates for small areas (domains) such as Federal Electoral Districts (FEDs) and Census Divisions (CDs), both of which number over 250 units across Canada, has increased in recent years. Since it was not possible to respect the boundaries of such areas in the design of the survey, various alternative small area estimation methodologies were evaluated. A sample dependent estimator was proposed as a combination of post-stratified and synthetic estimators, which relies entirely on the post-stratified estimator whenever the sample size in the domain is sufficient according to certain criteria, and which otherwise introduces a synthetic component whose relative weight depends on the deficiency of the sample in the domain. Based on study findings, it was recommended that the sample dependent approach be developed as a means of providing annual or multi-year average estimates for areas such as FEDs and CDs (Drew, Singh, and Choudhry 1982). Implementation and further research and developmental work is proceeding under Statistics Canada's Small Area Data Program.

### 7.4 Variance Estimation

The methodology for variance estimation for the redesigned sample will continue to be based on Keyfitz's (1957) method, although it will be further modified to the case of a two step final stage ratio estimation, i.e., to a single iteration in the raking ratio estimation procedure. As subsequent iterations exert only a very small influence on estimates, they are being ignored in variance estimation. Some further refinements of the current variance estimation procedure are under study, such as adopting clusters as replicates in SRUs, as opposed to the current practice of grouping clusters into two pseudo-replicates.

It should be noted that variance estimators given by Rao, Hartley, and Cochran (1962), and by Rao (1975) were evaluated as alternatives to the current method in SRUs, where the RHC design is followed (Choudhry, Lee, and Sida 1984). The current method and the alternatives were studied both with and without ratio adjustment. The current method without ratio adjustment was found to overestimate the variance for certain characteristics (e.g., 20% for employed), however with ratio estimation, biases were negligible. Estimated biases were also negligible for the alternatives. The principal advantage of the alternatives was that they were more stable. The current method was retained however, due to its simplicity and also because of the complications in estimating variances of change or averages under the alternative methods.

### 7.5 Composite Estimation

In the LFS, moderate to high month-to-month correlations exist for most characteristics due to the 5/6'th common sample. Different composite estimators were studied by Kumar and Lee (1983), which take advantage of these correlations by use of data from previous samples to improve the current month's estimates. Their studies focussed on a class of AK composite estimators studied recently by Huang and Ernst (1981) and others in the context of the U.S. Bureau of the Census' Current Population Survey.

Findings under the assumption that the ratio estimator is unbiased, were that from the perspective of mean square error, a compromise choice of the A and K weights yielded up to 5% gains for monthly estimates of level for unemployed and employed, and from 5% - 16% gains for corresponding month-to-month change estimates. A decision on implementation of composite estimation was delayed pending further studies on the impact on composite estimators of any changes in rotation group bias, stemming from modifications in non-response adjustment and ratio estimation procedures, and pending closer examination of its operational implications.

## 7.6 New Rounding and Release Policy

In the old LFS estimates of level were rounded to thousands and released if greater than 4 thousand. This policy was applied uniformly in all provinces for all estimates, with the intent that released data should have a coefficient of variation of 33.3% or less.

More rigorous, provincially based rounding and release criteria were developed for the redesigned sample to satisfy the conditions that the CV of unrounded estimates should be 33.3% or less, and that the rounding error should not exceed 20% of the standard error of the unrounded estimate. Findings were that release criteria could be dropped to 2 – 3 thousand, for all provinces except Quebec and Ontario, and that estimates for subprovincial areas should be rounded to hundreds instead of thousands (Kumar 1982).

## 8. SAMPLE REALLOCATION

Specific data reliability objectives of the redesign having particular emphasis on better data at subprovincial level are noted in Section 2. In addition to the general improvements in the data reliability levels through the refinements in the methods and procedures, it became necessary to consider reallocation of the sample within provinces to meet the reliability levels noted in objectives (ii), (iii) and (iv). Sample size increases were needed in 13 out of 66 Economic Regions, 6 out of 24 CMAs and 27 out of 42 non-CMA cities. An average 28% reduction in the CV's for unemployed was obtained for these cases. In addition, for the 30 ERs with old CV's in the range of 15 – 25%, the reallocations achieved an average 12% reduction in CV's. As a rule of thumb, under the redesigned sample, monthly data for ERs and CMAs and quarterly data for other cities will be based on minimum monthly sample sizes of 300 and 120 households per month respectively.

It is worth noting that the redesign objectives did not directly consider two important uses of LFS data by federal government departments. These are the use of 3 month moving average unemployment rates for subprovincial Unemployment Insurance (UI) Regions in determining the regionally variable number of weeks worked to qualify for UI benefits, and the use of 3 year average data for 180 – 200 areas consisting of individual or combined Census Divisions, for use in allocating federal funds to assist new industrial initiatives. However, the re-distributions of the sample will indirectly benefit both of these applications.

In determining sample size requirements to meet the objectives, average unemployment rates for the period 1980 – 82 were used, in view of medium term forecasts for sustained high unemployment during the 80's.

A general implication of these reallocations was the movement of a significant proportion of the sample from larger CMA's and Economic Regions to smaller ones. This had an adverse impact on the provincial and national estimates due to the departure from the usual proportional allocations. This decrease in reliability at higher levels was however more than compensated by the general increase in the reliability achieved through the structural improvements in the methods and procedures as a result of research investigations.

A study was also carried out using the cost-variance model suggested by Fellegi, Platek and Gray (1967) to arrive at optimum sampling rates in the NSR and SR areas. This resulted in a shift of sample from NSR to SR areas. This was most pronounced in Quebec and Ontario where the proportion of SR sample increased from .60 to .72, as compared with .78 of the frame, yielding gains for provincial estimates of unemployed equivalent to a 5% variance reduction, assuming a fixed sample size. In addition, this optimization helped in achieving objectives (ii) and (iv), and it benefitted the Survey of Consumer Finances and Rent Survey.

It is worth noting that the structural improvements in the design achieved through factors such as improved stratification, reduction in a stage of sampling in the NSR areas, incorporation of subprovincial controls in the estimation procedures, and more refined sample allocations resulted in better than current reliability levels for national and provincial estimates while meeting the objectives for the subprovincial data. This opened up the possibility of reducing



the overall sample of the LFS in order to release funds for the collection of data on certain other socio economic issues from time to time. The size of the redesigned sample was therefore fixed as 51,500 households per month down from 55,000. This overall reduction of 6 - 7% was achieved through a uniform reduction in all provinces with the exception of Prince Edward Island. In addition, per unit data collection costs will be reduced due to increased telephone interviewing.

## 9. IMPLICATION OF CHANGES ON LFS ASSOCIATED SURVEYS

Most of the household surveys conducted by Statistics Canada take advantage of the investment the LFS represents in terms of sample frame and design, data collection vehicle and processing systems to obtain data more quickly, at less cost and greater reliability than would be possible through independent surveys. The design and operations of these surveys are integrated with those of the LFS to varying degrees.

Most common are supplementary surveys consisting of additional questions to LFS respondent, which incur only incremental costs of the extra questions. Surveys which due to the sensitivity of the subject matter due to the length of the questionnaire, might impact on the LFS are not done as supplements. Typically such surveys have been conducted by LFS interviewers on a separate sample of households selected in the same areas as the LFS. Less closely integrated with the LFS are surveys which select different areas from the LFS, but which benefit from use of the LFS sample design and from control of overlap with the LFS sample.

As noted earlier the main orientation of the redesign program was towards the LFS, but efforts were also made to enhance flexibility of the vehicle for general applications. In this light changes being adopted for the LFS that will benefit its associated surveys are briefly highlighted below:

The sample reallocation resulting in a shift of sample from NSR to SR areas will be more robust for general applications and in particular will improve estimation of income and rent changes from the SCF and Rent Survey. Also, the adoption of a 300 household minimum sample size for CMAs will benefit these surveys for which CMA estimates are produced.

The general multi-variate stratification using 15 variables adopted (in both NSR and SR areas) will also represent an improvement for non-LFS applications over the current industry specific or simple geographic stratification.

Three changes will specifically benefit surveys using different sets of households: (i) the elimination of a stage of sampling in rural areas will considerably shorten the lead time to 7 months from 13 months in the old design, (ii) PSUs will be more geographically compact due to the adoption of explicit rural and urban strata, which will benefit smaller surveys where greater correspondence between PSUs and interviewers assignments is needed, and (iii) the flexibility introduced through the refinements in the sample stabilization program will allow selection of subsamples of virtually any size at the national, provincial or subprovincial levels for surveys using the LFS vehicle.

Finally the modification in the method of estimation introduced in the redesign, in the use of a household size distribution in the ratio adjustment as an interim measure (with eventual incorporation of demographic estimates of families by size as an additional dimension in the final staged raking ratio estimation procedure) will improve the consistency amongst family and individual related labour market statistics, and will improve family statistics on expenditure and income.

## 10. SUMMARY OF CHANGES AND FUTURE RESEARCH PLANS

Most of the research investigations for the post-1981 Censal redesign of the LFS have been completed, with the implementation of the new sampling design underway and certain aspects of research in estimation methodology still to continue. It is worth noting that many of the

investigations carried out during this program have confirmed the soundness of the past methods and procedures used in the LFS such as those of sampling two PSUs per stratum in the NSR areas, use of the RHC method and existing density factor (4 – 5 households per cluster) in the SR areas and continuation of the six month rotation pattern. However, several investigations have as well lead to improvements both in the redesign process and the new survey design; primarily due to change in emphasis on the data reliability objectives (as noted in Section 2), availability of better and easily accessible information and technological advances.

Improvements in the redesign process included the use of 1981 Census data in place of independently obtained field counts for updating the SR sample, the reduction of the clustering operation, and automation of stratification and PSU formation. Also cost savings will result from phasing-in much of the redesigned sample in an "on-line" fashion. Under this approach, the new sample will be introduced and the old sample will be dropped one rotation group at a time over a 6 month period, as compared with the traditional method of keeping the old sample at full strength for a 3 – 4 month period while building up the new sample (Mayda, Drew and Lindeyer 1984). Process cost savings as compared with the previous redesign are estimated at \$1.8 million (in 1983/84 dollars).

Principal improvements in the cost efficiency of the LFS survey design include the extension of telephone interviewing to months 2 – 6 in the sample in NSR areas, the adoption of an NSR design featuring explicit urban/rural stratification, elimination of a stage of sampling in rural areas, the general purpose stratification in both SR and NSR areas, the use of subprovincial population controls in the estimation procedure, and more refined sample allocation procedures. These improvements were sufficient to permit gains in the reliability of subprovincial data, while retaining the status quo for provincial level reliabilities, and while decreasing the overall sample size by 6 – 7%. Reliability gains averaged 14% for coefficients of variation of unemployed for the half of the Economic Regions and CMAs with poorest reliabilities under the old design, with for the most part, little or no change in remaining areas. Subprovincial gains for estimates of employed will be even greater. On the cost side, the sample size decrease, coupled with reduced costs due to the extension of telephone interviewing will result in estimated cost savings of \$.7 million per year (1983/84 dollars).

Following the completion of the sample redesign a principal focus of design related research and development for the LFS in coming years will be on investigation of a dual frame methodology, underwhich a portion of the sample would be converted to a telephone frame, using Random Digit Dialing (RDD) techniques. A multi-year program of RDD testing including research into implications of higher non-response to the RDD telephone sample, of research into dual frame estimation methodologies and of study of centralization versus decentralization of telephone interviewing is currently in the planning stages, as part of a newly established telephone survey development program (Hofmann, Drew, Catlin and Mayda 1984). Another design related initiative will be aimed at developing cost efficient means of intercensally updating the area sample in SR areas.

### ACKNOWLEDGEMENTS

The authors are grateful to the LFS Sample Redesign Committee for its support and guidance throughout the course of the redesign program, most notably to I.P. Fellegi, D.B. Petrie, G.J. Brackstone, R. Platek, I. Macredie, M. Levine, M. Brochu and F. Mayda. Special thanks are also due to all members of the methodology team engaged in both the research and implementation phases of the project, whose collective efforts have made possible the improvements described herein. The referee's helpful comments are also appreciated.

## REFERENCES

- CHOUDHRY, G.H., LEE, H. and DREW, J.D. (1984). Cost Variance Optimizations for the Canadian Labour Force Survey. Internal Technical Report (in preparation), Census and Household Survey Methods Division, Statistics Canada.
- CHOUDHRY, G.H., LEE, H. and SIDA, R. (1984). Variance Estimation for the Redesigned Labour Force Survey. Technical Report, Census and Household Survey Methods Division, Statistics Canada.
- DREW, J.D., CHOUDHRY G.H. and GRAY G.B. (1978). Some Methods for Updating Sample Survey Frames and Their Effects on Estimation. *Survey Methodology*, 4, 225-263.
- DREW, J.D., SINGH, M.P. and CHOUDHRY, G.H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 17-47.
- EARWAKER, S. and BÉLANGER, Y. (1981). Ratio Estimation at the Subprovincial Level for the Labour Force Survey. Technical Report, Census and Household Survey Methods Division, Statistics Canada.
- FELLEGI, I.P., GRAY, G.B. and PLATEK, R. (1967). The New Design of the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 62, 421-453.
- FOY, P., BÉLANGER, Y., DREW, J.D. and JONCAS, M. (1984). Multivariate Clustering Algorithm for Stratifications and its Application to the Canadian Labour Force Survey. Technical Report (in preparation), Census and Household Survey Methods Division, Statistics Canada.
- FRIEDMAN, H.P. and RUBIN, J. (1967). On some Invariant Criteria for Grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.
- GHANGURDE, P.D. (1984). Evaluation of LFS Non-Response Adjustment in Household Size Cells. Technical Report, Census and Household Survey Methods Division, Statistics Canada.
- GRAY, G.B. (1973). On Increasing the Sample Size (No. of PSUs). Technical Report, Census and Household Survey Methods Division, Statistics Canada.
- HARTLEY, H.O. and RAO, J.N.K. (1962). Sampling with Unequal Probabilities without Replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- HOFMANN, H., DREW, D., CATLIN, G., and MAYDA, F. (1984). A Proposal for a Telephone Survey Development Program. Internal Statistics Canada Report.
- HUANG, E. and ERNST, L. (1981). Comparison of an Alternate Estimator to the Current Composite Estimator in CPS. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 303-308.
- JUDKINS, D.R. and SINGH, R.P. (1981). Using Clustering Algorithms to Stratify Primary Sampling Units. *Proceedings of the Section on Survey Research Methods, American Statistical Association Meetings*, 274-284.
- KEYFITZ, N. (1951). Sampling with Probabilities Proportional to Size: Adjustment for Changes in the Probabilities. *Journal of the American Statistical Association*, 46, 105-109.
- KEYFITZ, N. (1957). Estimates of Sampling Variance where two Units are Selected from each Stratum. *Journal of the American Statistical Association*, 52, 503-510.
- KOSTANICH, D., JUDKIN, D., SINGH, R. and SCHANTZ, M. (1981). Modification of Friedman - Rubin's Clustering Algorithm for Use in Stratified PPS Sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association Meetings*.
- KUMAR, S. (1982). Investigation of the Labour Force Survey Rounding and Release Criteria. Technical Report, Census and Household Survey Methods Division, Statistics Canada.
- KUMAR, S. and LEE, H. (1983). Evaluation of Composite Estimation for the Canadian Labour Force Survey. *Survey Methodology*, 9, 178-201.
- LEMAITRE, G. (1983). Some Results from the Time and Cost Study. Technical Report, Census and Household Survey Methods Division, Statistics Canada.
- MACREDIE, I. (1983). Family Oriented Measures of Employment and Unemployment. Presented at OECD Working Party on Employment and Unemployment Statistics.



- MAYDA, F., DREW, D., and LINDEYER, J. (1984). Phase-in of the Redesign Labour Force Survey Sample. Technical Report, (in preparation), Census and Household Survey Methods Division, Statistics Canada.
- PAUL, E.C., and LAWES, M. (1982). Characteristics of Respondent and Non-Respondent Households in the Canadian Labour Force Survey. *Survey Methodology*, 8, 48-85.
- PLATEK, R. and SINGH, M.P. (1976). Methodology of the Canadian Labour Force Survey. Catalogue No. 71-526, Statistics Canada.
- PLATEK, R. and SINGH, M.P. (1977). A Strategy for Updating Continuous Surveys. *Metrika*, 25, 1-7.
- RAO, J.N.K. (1975). Unbiased Variance Estimation for Multi Stage Designs. *Sankhya*, Series C, 37, 133-139.
- RAO, J.N.K., HARTLEY, H.O. and COCHRAN, W.G. (1962). On a Simple Procedure of Unequal Probability Sampling without Replacement. *Journal of the Royal Statistical Society, Series B*, 24, 482-490.
- SINGH, M.P. and DREW, J.D. (1977). Sample Expansion in Self Representing Units of the Canadian Labour Force Survey. Technical Report, Census and Household Survey Methods Division, Statistics Canada.
- SINGH, M.P., and DREW, J.D. (1981a). Research Plans for the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association Meetings*.
- SINGH, M.P. and DREW, J.D. (1981b). Redesigning Continuous Surveys in a Changing Environment. *Survey Methodology*, 7, 44-73.
- VERMA, R.B.P., BASAVARAJAPPA, K.G. and BENDER, R.K. (1983). The Regression Estimates of Population for Sub-Provincial Areas in Canada. *Survey Methodology*, 9, 219-240.

## Analysis of Categorical Data from Surveys with Complex Designs: Some Canadian Experiences<sup>1</sup>

D.A. Binder, M. Gratton, M.A. Hidirolou,  
S. Kumar and J.N.K. Rao<sup>2</sup>

### ABSTRACT

Goodness of fit tests, tests for independence in a two-way contingency table, log-linear models and logistic regression models are investigated in the context of samples which are obtained from complex survey designs. Suggested approximations to the null distributions are reviewed and some examples from the Canada Health Survey and Canadian Labour Force Survey are given. Software implementation for using these methods is briefly discussed.

**KEYWORDS:**  $\chi^2$  statistic; Wald Statistics; Goodness of fit; Independence in two-way tables; Log-linear and logistic regression model.

### 1. INTRODUCTION

A sketch of the historical development of modern categorical data analysis has been given in the excellent review paper by Imrey, Koch and Stokes (1981). These techniques, applied in the context of random samples derived as independent selections from a common distribution function, are not directly applicable to survey samples collected using complex survey designs.

Koch *et al* (1975), Shuster and Downing (1976), developed asymptotically valid methods, based on the Wald statistic that take the survey design into account, but requiring access to the micro-data file or at least the full estimated covariance matrix of cell estimates. Cohen (1976) and Altham (1976) proposed a simple model for clustering and showed that the generalized Wald statistic for goodness of fit is a multiple of  $\chi^2$ , when the model holds. Brier (1978) considered a similar model, but studied general hypotheses on cell probabilities, and proved that a multiple of the corresponding Pearson statistic is asymptotically distributed as a  $\chi^2$  random variable, when the model holds. Fellegi (1980) deflated the  $\chi^2$  using a correction factor based on the mean of the estimated design effects. Fay (1985) developed jackknife  $\chi^2$  and  $G^2$  statistics, also taking the design into account, but requiring the cell estimated at the primary sampling unit level. Rao and Scott (1981) developed a correction to  $\chi^2$  (or  $G^2$ ) based on the Bartlettwaite to approximation to the asymptotic distribution of  $\chi^2$ , requiring the full estimated covariance matrix.

In this paper, we discuss the problems of fitting models and testing hypotheses with categorical data resulting from complex designs. For data collected using complex designs, some adjustments to the classical methods described by Imrey, Koch and Stokes (1981) are necessary in order to make valid inferences. If the published tables are provided along with the cell and marginal design

---

This paper is a revised and expanded version of that presented at the Seminar on Recent Developments in the Analysis of Large Scale Data Sets sponsored by Statistical Office of European Communities, November 16-18, 1983, Luxembourg.

D.A. Binder, Institutional & Agriculture Survey Methods Division, M. Gratton, EDP Planning and Support Division, M.A. Hidirolou, Business Survey Methods Division, S. Kumar, Census & Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, and J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, CANADA.

effects, some of the approximations to the null distributions of our test statistics can be obtained, without having access to the complete micro-data file. On the other hand, for applications where the complete micro-data file is available, alternative approaches will be described.

For illustrative purposes, Section 2 begins with the standard goodness of fit problem. This discussion is then extended in Section 3 to tests of independence in a two-way contingency table. This leads to a general discussion of log-linear models in Section 4. Logistic regression models are described in Section 5. In Section 6 we summarize the existing situation with respect to software development for these methods at Statistics Canada. In Section 7, we discuss the appropriateness of these methods. Numerical examples are taken primarily from the Canada Health Survey. An application from the Canadian Labour Survey is given in Section 5.

## 2. GOODNESS OF FIT

### 2.1 Multinomial Sampling

Suppose we select  $n$  independent and identically distributed observations  $Y_1, \dots, Y_n$  from a discrete distribution with  $k$  categories, where  $\Pr(Y = i) = \pi_i$ ;  $\sum_{i=1}^k \pi_i = 1$ . We observe the random vector  $\underline{n} = (n_1, \dots, n_{k-1})^T$ , which has a multinomial distribution. Our estimate of  $\pi = (\pi_1, \dots, \pi_{k-1})^T$  is given by  $\underline{p} = \underline{n}/n$ . This estimate is unbiased and has covariance matrix given by  $V\{\underline{p}\} = (\underline{D}_\pi - \pi\pi^T)/n = \underline{P}/n$ , where  $\underline{D}_\pi = \text{diag}\{\pi_1, \dots, \pi_{k-1}\}$ . Note that  $\underline{P}^{-1} = \underline{D}_\pi^{-1} + (\mathbf{1}\mathbf{1}^T/\pi_k)$ . Asymptotically,  $n^{1/2}(\underline{p} - \pi) \rightarrow N(\underline{0}, \underline{P})$ . For a given  $\pi_o$ , the goodness of fit problem is to test the hypothesis.

$$H_o: \pi = \pi_o,$$

against the alternative

$$H_1: \pi \neq \pi_o. \quad (2.1)$$

Letting  $\underline{P}_o$  represent  $\underline{P}$  evaluated at  $\pi_o$ , the Wald statistic for this test is

$$\begin{aligned} W_1 &= n(\underline{p} - \pi_o)^T \underline{P}_o^{-1} (\underline{p} - \pi_o) \\ &= n \sum_{i=1}^K \{(p_i - \pi_{io})^2 / \pi_{io}\}, \end{aligned}$$

which is the familiar Pearson chisquare test. Under  $H_o$  this is asymptotically  $\chi_{k-1}^2$ . The likelihood ratio test for this problem is given by

$$LR_1 = 2n \sum_{i=1}^k p_i \log(p_i / \pi_{io}).$$

Since  $2p_i \log(p_i / \pi_{io})$  is asymptotically equivalent to  $2(p_i - \pi_{io}) + (p_i - \pi_{io})^2 / \pi_{io}$  under  $H_o$ , we see that the likelihood ratio test is asymptotically equivalent to the Pearson chisquare statistic under  $H_o$ .

Another possible test for this hypothesis is derived by defining the vector of logs,  $\underline{\mu}_o = \log \pi_o$  and  $\hat{\underline{\mu}} = \log \underline{p}$ . Now under the null hypotheses  $\hat{\underline{\mu}} - \underline{\mu}_o$  is asymptotically equivalent to  $\underline{D}_{\pi_o}^{-1}(\underline{p} - \pi_o)$ . Therefore,  $n^{1/2}(\hat{\underline{\mu}} - \underline{\mu}_o) \rightarrow N(\underline{0}, \underline{D}_{\pi_o}^{-1} - \mathbf{1}\mathbf{1}^T)$  under  $H_o$  and the Wald statistic is

$$\begin{aligned} W_2 &= (\hat{\underline{\mu}} - \underline{\mu}_o)^T [\underline{D}_{\pi_o} + (\pi_o \pi_o^T / \pi_{ko})] (\hat{\underline{\mu}} - \underline{\mu}_o) \\ &= \sum_{i=1}^k \pi_{io} (\hat{\mu}_i - \mu_{io})^2, \end{aligned}$$

where  $\mu_{ko} = \log \pi_{ko}$  and  $\hat{\mu}_k = \log p_k$ .

This approximation is obtained by noting that under  $H_o$

$$\begin{aligned} \pi_{ko}(\hat{\mu}_k - \mu_{ko}) &\doteq p_k - \pi_{ko} \\ &= - (p - \pi_o)^T \underline{1} - (\underline{\mu} - \underline{\mu}_o)^T \underline{\pi}_o. \end{aligned}$$

Note that  $W_2$  is also asymptotically equivalent to the Pearson chisquare test under  $H_o$ .

2.2 Other Sampling Schemes

These results for  $W_1$ ,  $W_2$  and  $LR_1$  are well-known. The question of interest to us here is the implication of the more general assumption that  $n^{1/2}(p - \pi) \rightarrow N(0, V)$ , where  $V$  is not necessarily equal to  $P$ . Here  $p$  is a survey estimate of  $\pi$  and may depend on sampling weights and other adjustment factors. This situation often arises in sampling under a complex sample design. We assume that  $\hat{V}$  is a consistent estimate of  $V$ . There are two approaches which we shall consider here. The first is to construct the appropriate Wald statistic for the given sample design. This would be

$$W_3 = n(p - \pi_o)^T \hat{V}^{-1} (p - \pi_o),$$

where the rank of  $\hat{V}$  is  $k-1$  so that  $W_3$  is asymptotically  $\chi^2_{k-1}$  under  $H_o$ .

An alternative approach would be to use  $W_1$ ,  $W_2$  or  $LR_1$  directly as a test statistic. Now from multivariate normal theory, we know that the distribution of  $n(p - \pi_o)^T P_o^{-1} (p - \pi_o)$  is that of  $\sum \delta_i Z_i^2$ , where  $\{Z_i^2\}$  are independent  $\chi^2_1$  random variable and  $\delta = (\delta_1, \dots, \delta_{k-1})^T$  are the eigenvalues of  $P_o^{-1} V$ ; see Johnson and Kotz (1970, pg. 150). This result was shown by Rao and Scott (1981), who call the  $\delta_i$ 's generalized design effects. We note that for  $k = 2$ , we have  $\delta = n\sigma_p^2 / \{\pi_o(1 - \pi_o)\}$ , where  $\sigma_p^2 = V\{p\}$ . This is the usual design effect for  $p$  under  $H_o$ .

2.3 Approximations

Now, in general, the distribution function for linear combinations of  $\chi^2_1$  random variables is cumbersome, although their moments are easily obtained. Rao and Scott (1981) have suggested two approximations to obtain the significance levels. The first is to approximate the distribution as being proportional to a  $\chi^2_{k-1}$  random variable, the proportionality constant being determined by equating the mean of the approximating distribution to that of the theoretical distribution. This results in the approximation

$$\sum_{i=1}^{k-1} \delta_i Z_i^2 \doteq \left\{ \sum_{i=1}^{k-1} \delta_i / (k-1) \right\} \chi^2_{k-1} \tag{2.2}$$

Now,

$$\begin{aligned} \sum \delta_i &= \text{tr}(P_o^{-1} V) \\ &= \sum_{i=1}^k v_{ii} / \pi_{io} \\ &= \sum_{i=1}^k d_i (1 - \pi_{io}), \end{aligned}$$

which depends only on the cell design effects  $\{d_i\}$ , where  $v_{ii}$  is the  $i$ -th diagonal element of  $V$  and  $d_i = v_{ii} / [\pi_{io}(1 - \pi_{io})]$ . This approximation is particularly convenient when the full covariance matrix is not known, but the cell design effects are given. This is often the case for official published data.

**Table 1**  
Age Distribution Among Those Consuming 1-6 Drinks Per Week.  
Census Age Distribution for Canada (1978-9)

Census Distribution	Age							Total
	15-19	20-24	25-34	35-44	45-54	55-64	65 +	
	.133	.127	.218	.152	.140	.115	.115	1.000
Distribution of those consuming 1-6 drinks/week	.117	.150	.264	.175	.148	.093	.053	1.000
Design Effect	1.4	1.2	2.2	1.1	0.6	1.1	1.0	

### Example 1

For the Canada Health Survey (1978-9), a stratified multi-stage household survey, data was derived for the age distribution among those consuming one to six drinks per week, based on a sample of 5,204 persons, aged 15 years and over. A description of the survey may be found in "The Health of Canadians" (Statistics Canada Catalogue No. 82-538).

The data, taken from Hidiroglou and Rao (1981), are presented in Table 1. The raw value for  $W_1$  is 298. This is reduced to 248 by taking the approximation given by (2.2). For these data, the post-stratification adjustments for age and sex lead to small design effects.

A second approximation to the distribution of  $\sum \delta_i Z_i^2$ , suggested by Rao and Scott (1981), is the Satterthwaite (1946) approximation:  $\sum \delta_i Z_i^2 \approx a\chi_{\nu}^2$ . To obtain  $a$  and  $\nu$ , it is necessary to compute

$$\begin{aligned}\Sigma \delta_i^2 &= \text{tr}\{(\mathbf{P}_o^{-1}\hat{\mathbf{V}})^2\} \\ &= \sum_{i=1}^k \sum_{j=1}^k v_{ij}^2 / (\pi_{i0}\pi_{j0}).\end{aligned}$$

However, this depends on all the terms of the matrix  $\hat{\mathbf{V}}$ . The important point, though, is that some adjustment to the multinomial test statistic is necessary to obtain the appropriate significance level.

An alternative approximation, suggested by Fellegi (1980), is to divide the statistic  $n(p - \pi_o)^T \mathbf{P}_o^{-1}(p - \pi_o)$  by the average design effect,  $\bar{d}$ , instead of the weighted average given in (2.2). The effect of this on the data in Table 1 is that the adjusted chisquare value is 243, which is comparable to Rao and Scott's (1981) approximation.

## 3. TESTS OF INDEPENDENCE IN A TWO-WAY TABLE

### 3.1 Multinomial Sampling

We now suppose that the categories of the multinomial distribution can be cross-classified into an  $r \times c$  table, where for the bivariate observation  $(Y_1, Y_2)$  we have  $\Pr(Y_1 = i, Y_2 = j) = \pi_{ij}$ ;  $\sum_{i=1}^r \sum_{j=1}^c \pi_{ij} = 1$ . We denote  $\pi_{i+} = \sum_{j=1}^c \pi_{ij}$  and  $\pi_{+j} = \sum_{i=1}^r \pi_{ij}$ . We denote  $\pi = (\pi_{11}, \dots, \pi_{1c}, \dots, \pi_{r1}, \dots, \pi_{rc})^T$ ,  $\pi_R = (\pi_{1+}, \dots, \pi_{(r-1)+})^T$ ,  $\pi_C = (\pi_{+1}, \dots, \pi_{+(c-1)})^T$ ,  $\mathbf{P}_R = \mathbf{D}_{\pi_R} - \pi_R \pi_R^T$ ,  $\mathbf{P}_C = \mathbf{D}_{\pi_C} - \pi_C \pi_C^T$ . We observe the random vector  $\underline{n}$  from the multinomial distribution, where  $E\{\underline{n}\} = n\pi$ . We let  $\underline{p} = \underline{n}/n$ ,  $p_{i+} = \sum_j p_{ij}$ , and  $p_{+j} = \sum_i p_{ij}$ .



We wish to test the hypothesis of independence

$$H_0: \pi_{ij} - \pi_{i+}\pi_{+j} = 0 \text{ for } 1 \leq i \leq r-1; 1 \leq j \leq c-1,$$

against the alternative

$$H_1: \pi_{ij} - \pi_{i+}\pi_{+j} \neq 0 \text{ for some } (i, j).$$

If we construct  $h_{ij} = p_{ij} - p_{i+}p_{+j}$ , for  $1 \leq i \leq r-1$  and  $1 \leq j \leq c-1$ , then under multinomial sampling under  $H_0$ , the asymptotic covariance matrix for  $\underline{h} = (h_{11}, \dots, h_{1,c-1}, \dots, h_{r-1,c-1})^T$  is  $\underline{P}_R \otimes \underline{P}_C$ , where  $\otimes$  denotes the direct matrix product operation. Hence, the Wald statistic under  $H_0$  becomes

$$\begin{aligned} W_4 &= \underline{h}^T (\hat{\underline{P}}_C^{-1} \otimes \hat{\underline{P}}_R^{-1}) \underline{h} \\ &= \sum_{i=1}^r \sum_{j=1}^c (p_{ij} - p_{i+}p_{+j})^2 / (p_{i+}p_{+j}), \end{aligned}$$

the familiar chisquare test with  $(r-1)(c-1)$  degrees of freedom.

Another test, which is asymptotically equivalent to  $W_4$  under  $H_0$ , is the likelihood ratio test given by

$$LR_2 = 2n \left[ \sum_{i=1}^r \sum_{j=1}^c p_{ij} \log p_{ij} - \sum_{i=1}^r p_{i+} \log p_{i+} - \sum_{j=1}^c p_{+j} \log p_{+j} \right].$$

An alternative approach for this problem, which is a special case of the methods described by Grizzle, Starmer and Koch (1969) is to consider a Wald statistic based on

$$\{f_{ij} = \log p_{ij} - \log p_{i+} - \log p_{+j}; \text{ for } 1 \leq i \leq r-1, \text{ and } 1 \leq j \leq c-1\}.$$

The asymptotic covariance matrix for  $\underline{f} = (f_{11}, \dots, f_{1,c-1}, \dots, f_{r-1,c-1})^T$  is  $(\underline{D}_{\pi_R}^{-1} - \underline{1}\underline{1}^T)$  ( $\underline{D}_{\pi_C}^{-1} - \underline{1}\underline{1}^T$ ). Therefore the Wald statistic becomes

$$W_5 = \underline{f}^T \left[ (\hat{\underline{D}}_{\pi_R} + \frac{\underline{\pi}_R \underline{\pi}_R^T}{p_{r+}}) \otimes (\hat{\underline{D}}_{\pi_C} + \frac{\underline{\pi}_C \underline{\pi}_C^T}{p_{+c}}) \right] \underline{f}$$

Now under  $H_0$  we note that  $f_{ij}$  is asymptotically equivalent to

$$\frac{p_{ij}}{\pi_{i+}\pi_{+j}} - \frac{p_{i+}}{\pi_{i+}} - \frac{p_{+j}}{\pi_{+j}} + 1,$$

so that  $\sum_{i=1}^r \pi_{i+} f_{ij} = \sum_{j=1}^c \pi_{+j} f_{ij} = 0$ . Using this approximation  $W_5$  becomes

$$W_5' = \sum_{i=1}^r \sum_{j=1}^c p_{i+} p_{+j} f_{ij}^2.$$

It should be noted that under  $H_0$ , the statistics  $W_4$ ,  $LR_2$  and  $W_5'$  are all asymptotically equivalent to

$$\sum_{i=1}^r \sum_{j=1}^c \frac{(p_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}} = \sum_{i=1}^r \frac{(p_{i+} - \pi_{i+})^2}{\pi_{i+}} = \sum_{j=1}^c \frac{(p_{+j} - \pi_{+j})^2}{\pi_{+j}} \quad (3.1)$$

This result will prove useful in Section 3.3.



### 3.2 Other Sampling Schemes

Now, relaxing the assumption that  $\underline{n}$  is multinomial, we assume instead that  $n^{1/2}(\underline{p} - \underline{\pi}) \rightarrow N(0, \underline{V})$  where  $\underline{p}$  is a survey estimate which may depend on sampling weights and other adjustment factors. For this case, Shuster and Downing (1976) and Fellegi (1980) suggest that we construct the Wald statistic based on  $\{h_{ij} = p_{ij} - p_{i+}p_{+j}\}$ . If we let  $\underline{J}_a$  be the  $(a-1) \times a$  matrix given by

$$\underline{J}_a = [\underline{I} \mid 0] \quad (3.2)$$

$$\text{and let } \hat{H} = (\underline{J}_r - \underline{\pi}_R \underline{1}^T) \otimes (\underline{J}_c - \underline{\pi}_C \underline{1}^T) - (\underline{\pi}_R \underline{1}^T \otimes \underline{\pi}_C \underline{1}^T)$$

then the Wald statistics is

$$W_6 = \underline{h}^T (\hat{H} \hat{V} \hat{H}^T)^{-1} \underline{h},$$

which under  $H_0$  is asymptotically  $\chi^2_{(r-1)(c-1)}$ .

Alternatively, we could construct a Wald statistic based on  $\{f_{ij} = \log p_{ij} - \log p_{i+} - \log p_{+j}\}$ . This is a special case of the log-linear model approach to be discussed in Section 4. We define  $(r-1) \times r$  and  $(c-1) \times c$  matrices as follows:

$$\hat{\underline{E}}_R = \left[ \hat{\underline{D}}_{\pi_R}^{-1} \mid 0 \right], \quad \hat{\underline{E}}_C = \left[ \hat{\underline{D}}_{\pi_C}^{-1} \mid 0 \right].$$

$$\text{We let } \hat{F} = (\hat{\underline{E}}_R - \hat{\underline{E}}_R \underline{1} \underline{1}^T) \otimes (\hat{\underline{E}}_C - \hat{\underline{E}}_C \underline{1} \underline{1}^T) - (\hat{\underline{E}}_R \underline{1} \underline{1}^T \otimes \hat{\underline{E}}_C \underline{1} \underline{1}^T).$$

The appropriate Wald statistics is

$$W_7 = \underline{f}^T (\hat{F} \hat{V} \hat{F}^T)^{-1} \underline{f}.$$

Now, analogously to the goodness of fit problem in Section 2, Rao and Scott (1981) have considered null distributions of the test statistics based on  $W_4$ ,  $LR_2$  and  $W_5$ , which are all asymptotically equivalent to the null distribution of (3.1). We see, therefore, that the null distribution is the same as  $\sum_{i=1}^{(r-1)(c-1)} \delta_i Z_i^2$ , where  $\{Z_i^2\}$  are independent  $\chi_1^2$  and the  $\delta_i$ 's are the eigenvalues of

$$(\underline{P}_R^{-1} \otimes \underline{P}_C^{-1})(\underline{H} \underline{V} \underline{H}^T).$$

Cowan and Binder (1978) investigated the properties of the eigenvalue from a simple two-stage self-weighting design for a  $2 \times 2$  table. They found that the eigenvalue increases as the degree of independence of the cell proportions within the primary sampling units decreased.

3.3 Approximations

An approximation for the distribution of  $\sum \delta_i Z_i^2$  is

$$\sum \delta_i Z_i^2 \approx \frac{\sum \delta_i}{(r-1)(c-1)} \chi^2_{(r-1)(c-1)},$$

as in (2.2). Since the statistic is asymptotically equivalent to (3.1) under  $H_0$ , by computing the mean of (3.1) we obtain

$$\sum \delta_i = \sum_{i=1}^r \sum_{j=1}^c d_{ij} (1 - \pi_{i+} \pi_{+j}) - \sum_{i=1}^r d_i^{(r)} (1 - \pi_{i+}) - \sum_{j=1}^c d_j^{(c)} (1 - \pi_{+j}),$$

where  $d_{ij}$  is the cell design effect;  $d_i^{(r)}$  and  $d_j^{(c)}$  are the row and column margin design effects, respectively. This particularly simple expression was obtained by Rao and Scott (1983). Fellegi (1980) suggested an alternative approximation as:

$$(\sum_{i=1}^r \sum_{j=1}^c d_{ij} / rc) \chi^2_{(r-1)(c-1)}$$

Example 2

In Table 2, we give a  $4 \times 2$  table from the Canada Health Survey, which cross-classifies drug use (four categories; 0, 1, 2, 3+ drug classes in a 2-day period) and sex (male, female). Here  $r = 31,668$ .

The raw value for  $W_4$  is 774. Rao and Scott's (1981) adjustment reduces this to 437. Fellegi's (1980) adjustment reduces this to 327. The Wald statistics,  $W_6$ , is 538. Hidiroglou and Rao (1981) found that the Rao and Scott (1981) approximation performs quite well relative to the atterthwaite (1946) approximation which is based on the complete covariance matrix.

LOG-LINEAR MODELS

1 Multinomial Sampling

We now extend the results of the previous section to more general cross-classifications of the multinomial distribution. The standard results for these models are given in Bishop, Fienberg

Table 2  
Variety of Drugs Taken by Sex for Canada (1978-79)

Sex		Number of Drug Varieties				Total
		0	1	2	3±	
Male	Proportion	0.293	0.134	0.048	0.021	0.496
	Design Effect	1.56	3.37	1.15	1.38	0.00*
Female	Proportion	0.228	0.159	0.072	0.045	0.504
	Design Effect	3.59	3.13	2.85	1.96	0.00*
Total	Proportion	0.521	0.293	0.120	0.066	1.000
	Design Effect	6.03	6.46	1.65	2.57	

\* Because of age-sex post-stratification, these design effects are zero.

and Holland (1975) and Fienberg (1980). We have  $\pi = (\pi_1, \dots, \pi_k)^T$  is a vector of cell proportions;  $\sum_{i=1}^k \pi_k = 1$ . We observe  $\underline{n} = (n_1, \dots, n_k)^T$ , the counts in each cell from a random sample, so that  $\underline{n}$  has a multinomial distribution ( $\sum n_i = n$ ). We let  $\underline{p} = \underline{n}/n$  and define

$$\underline{\mu} = \log \pi.$$

The log-linear model assumes that for a parameter vector  $\underline{\theta} = (\theta_1, \dots, \theta_t)^T$ , we have

$$\underline{\mu}(\underline{\theta}) = u(\underline{\theta})\underline{1} + \underline{X}\underline{\theta},$$

where  $\underline{X}$  is a known  $k \times t$  matrix of full rank and  $\underline{X}^T \underline{1} = \underline{0}$ . Note that  $t \leq k-1$ . If  $t = k-1$ , we have the saturated model.

The maximum likelihood estimate for  $\underline{\theta}$  is given by solving

$$\underline{X}^T(\underline{p} - \hat{\underline{\pi}}) = \underline{0}, \quad (4.1)$$

where  $\hat{\underline{\pi}} = \pi(\hat{\underline{\theta}})$ . Now, asymptotically we have

$$\hat{\underline{\pi}} - \underline{\pi} \doteq \underline{P}\underline{X}(\hat{\underline{\theta}} - \underline{\theta}),$$

where  $\underline{P} = \underline{D}_{\underline{\pi}} - \underline{\pi}\underline{\pi}^T$ . From (4.1), we then obtain

$$\hat{\underline{\theta}} - \underline{\theta} \doteq (\underline{X}^T \underline{P} \underline{X})^{-1} \underline{X}^T (\underline{p} - \underline{\pi})$$

and

$$\hat{\underline{\pi}} - \underline{\pi} \doteq \underline{P}\underline{X}(\underline{X}^T \underline{P} \underline{X})^{-1} \underline{X}^T (\underline{p} - \underline{\pi}).$$

Since  $n^{1/2}(\underline{p} - \underline{\pi}) \rightarrow N(\underline{0}, \underline{P})$  we obtain

$$n^{1/2}(\hat{\underline{\theta}} - \underline{\theta}) \rightarrow N[\underline{0}, (\underline{X}^T \underline{P} \underline{X})^{-1}]$$

$$n^{1/2}(\hat{\underline{\pi}} - \underline{\pi}) \rightarrow N[\underline{0}, \underline{P}\underline{X}(\underline{X}^T \underline{P} \underline{X})^{-1} \underline{X}^T \underline{P}].$$

Suppose now that the linear expression  $\underline{X}\underline{\theta}$  can be decomposed as  $\underline{X}_1 \underline{\theta}_1 + \underline{X}_2 \underline{\theta}_2$  where  $\underline{X}$  and  $\underline{X}_2$  are full rank,  $\underline{X}_1$  is  $k \times r$ ,  $\underline{X}_2$  is  $k \times s$ ,  $\underline{\theta}_1$  is  $r \times 1$  and  $\underline{\theta}_2$  is  $s \times 1$ , where  $r + s = t$ .

We consider the problem of testing

$$H_0: \underline{\theta}_2 = \underline{0},$$

against the alternative

$$H_1: \underline{\theta}_2 \neq \underline{0}.$$

We use  $\underline{\theta}_1, \underline{\theta}_2, \underline{\pi}$ , etc. to denote the estimates under the full model  $H_1$ . Alternatively, we let  $\hat{\underline{\theta}}_1, \hat{\underline{\theta}}_2$  to denote estimates under  $H_0$ .

Now,

$$n^{1/2}(\underline{\theta}_2 - \underline{\theta}_2) \rightarrow N[\underline{0}, (\bar{\underline{X}}_2^T \underline{P} \bar{\underline{X}}_2)^{-1}]$$

where

$$\bar{\underline{X}}_2 = [\underline{I} - \underline{X}_1(\underline{X}_1^T \underline{P} \underline{X}_1)^{-1} \underline{X}_1^T \underline{P}] \underline{X}_2 \quad (4.2)$$

so that the Wald statistic is

$$W_8 = n \hat{\theta}_2^T \bar{X}_2^{-1} \hat{P} \bar{X}_2 \theta_2.$$

Under  $H_0$ , this is asymptotically equivalent to the Pearson chisquare statistic

$$n (\hat{\pi} - \hat{\pi})^T \hat{D}_\pi^{-1} (\hat{\pi} - \hat{\pi}),$$

or the likelihood ratio test

$$LR_3 = 2n \sum_{i=1}^k p_i \log(\hat{n}_i / \hat{\pi}_i).$$

Under  $H_0$ , these statistics are asymptotically  $\chi^2_s$ .

4.2 Other Sampling Schemes

We still assume that the cell proportions,  $\pi$ , satisfy  $\mu = \log \pi \ u(\theta_1, \theta_2) \mathbf{1} + X_1 \theta_1 + X_2 \theta_2$  but we now have  $n^{1/2}(p - \pi) \rightarrow N(0, V)$ , where  $p$  is a survey estimate.

Rao and Scott (1983) suggest the following Wald statistic for testing  $\theta_2 = 0$ . We let  $C$  be any  $k \times s$  matrix with  $C^T X_1 = 0$ ,  $C^T \mathbf{1} = 0$  and  $C^T X_2$  nonsingular. For example if  $X_1^T X_2 = 0$  then  $C = X_2$  is convenient. Now the hypothesis is equivalent to  $C^T \mu = 0$ . We have

$$\begin{aligned} C^T(\hat{\mu} - \mu) &\doteq C^T \hat{D}_\pi^{-1} (\hat{\pi} - \pi) \\ &\doteq C^T X (X^T P X)^{-1} X^T (p - \pi), \end{aligned}$$

where  $\pi$  is obtained from (4.1), based on the survey estimate,  $p$ .

We therefore have the Wald statistics

$$W_9 = n \hat{\mu}^T C [C^T X (X^T \hat{P} X)^{-1} (X^T \hat{V} X) (X^T \hat{P} X)^{-1} X^T C]^{-1} C^T \hat{\mu}.$$

Similar results were also given in Binder (1983). If under  $H_1$ , the model is saturated ( $r+s = k-1$ ), then  $p = \pi$  and we obtain

$$W_9 = n \hat{\mu}^T C [C^T \hat{D}_\pi^{-1} \hat{V} \hat{D}_\pi^{-1} C]^{-1} C^T \hat{\mu}.$$

Rao and Scott (1984) show that if we use  $\hat{P}$  instead of  $\hat{V}$  in  $W_9$  then these are asymptotically equivalent to the likelihood ratio or Pearson  $\chi^2$  test statistics. They also show that the likelihood ratio test statistics is distributed as  $\sum_i \delta_i Z_i^2$  under  $H_0$ , where  $\{Z_i^2\}$  are independent  $\chi^2_1$  and  $\{\delta_i\}$  are the eigenvalues of

$$(\bar{X}_2^T P \bar{X}_2)^{-1} (\bar{X}_2^T V \bar{X}_2), \tag{4.3}$$

for  $\bar{X}_2$  defined in (4.2).

4.3 Approximations

As before, we approximate the null distribution

$$\sum_{i=1}^s \delta_i Z_i^2 \approx \left( \frac{\sum \delta_i}{s} \right) \chi^2_s.$$

This involves computing the trace of (4.3). Rao and Scott (1984) show that if the model admits explicit solutions for both  $\hat{\pi}$  and  $\hat{\hat{\pi}}$ , then the approximation depends on the matrix  $V$  only through cell design effects and marginal design effects. This observation is particularly convenient when only the estimated design effects for the cell proportions and margins are available, as is often the case for published tables.

### Example 3

Hidioglou and Rao (1983) considered all direct estimates from the three-way table: Drug use (5 categories: 0, 1, 2, 3, 4+ drug classes in a 2 day period)  $\times$  Age (4 categories; 0-14, 15-44, 45-64, 65+)  $\times$  Sex (male, female), taken from the Canada health Survey. We give the results for testing whether Age and Sex are independent in each drug category ( $n = 31,668$ ). This is equivalent to the hypothesis

$$H_0: \pi_{ijk} = \pi_{ij+} \frac{\pi_{i+k}}{\pi_{i++}}.$$

Using Bishop, Fienberg and Holland's (1975) notation, where  $\log \pi_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$ , the null hypothesis is equivalent to

$$H_0: u_{23(jk)} = u_{123(ijk)} = 0 \quad \text{for all } (i, j, k).$$

The raw chisquare value is 23 based on 15 degrees of freedom. The average eigenvalue is 1.39, so that the approximation reduces the chisquare value to 16. Whereas the unadjusted chisquare value would lead the analyst to reject the hypothesis at the 10% level, the approximation indicates that  $h_0$  cannot be rejected even at the 30% level.

## 5. LOGISTIC REGRESSION MODELS

### 5.1 Multinomial Sampling

We now consider a logistic regression model for the conditional distribution of a binary response variable  $y$  given the vector  $x$  of independent variables. In particular, this conditional distribution is

$$\Pr(y_i | x_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i},$$

where  $y_i \in \{0, 1\}$ .

For the logistic regression model, we have

$$\log \left\{ \frac{\pi(x_i)}{1 - \pi(x_i)} \right\} = x_i^T \theta,$$

where  $\theta$  is an unknown vector of parameters.

We note that if  $x_i$  is a categorical vector of 0's and 1's, this is a special case of a log-linear model as described in Section 4. Here we allow  $x_i$  to be arbitrary. The extension to the case of  $k$ -categories for the  $y$ -variable is straight-forward, it is also possible to generalize the model to

$$\log \left\{ \frac{\pi(x_i)}{1 - \pi(x_i)} \right\} = f(x_i^T \theta),$$

for a known function  $f(\cdot)$ , but we do not discuss this here.

Now, the maximum likelihood estimate for  $\theta$  is given by

$$X^T(y - \hat{\pi}) = 0$$

where  $y = (y_1, \dots, y_n)^T$ ,  $\hat{\pi} = [\hat{\pi}(x_1), \dots, \hat{\pi}(x_n)]^T$  and  $X = [x_1 \dots x_n]^T$ .

Under suitable regularity conditions, we have

$$n^{1/2}(\hat{\theta} - \theta) \rightarrow N[0, n(X^T \Lambda X)^{-1}], \text{ where } \Lambda = D_{\pi}(\underline{I} - D_{\pi}).$$

If we have  $X\theta = X_1\theta_1 + X_2\theta_2$  and consider testing the hypothesis

$$H_0: \theta_2 = 0$$

$$H_1: \theta_2 \neq 0,$$

we obtain the Wald statistic

$$W_{10} = n \hat{\theta}_2^T (\hat{X}_2^T \Delta \hat{X}_2)^{-1} \hat{\theta}_2$$

where

$$\hat{X}_2 = [I - X_1(X_1^T \Delta X_1)^{-1} X_1^T \Delta] X_2.$$

The likelihood ratio test here is

$$LR_4 = 2 \sum_{i=1}^n \left[ y_i \log \left( \frac{\hat{\pi}_i}{\hat{\pi}_i} \right) + (1 - y_i) \log \left\{ \frac{(1 - \hat{\pi}_i)}{(1 - \hat{\pi}_i)} \right\} \right]$$

which is asymptotically equivalent to  $W_{10}$  under  $H_0$ .

## 5.2 Other Sampling Schemes

Suppose now that  $n^{-1/2} X^T(y - \pi) \rightarrow N(0, V)$  and that  $\hat{V}$  is a consistent estimator of  $V$ . Here  $y$  is not necessarily a vector of 0's and 1's, but may in fact depend on the sampling weights and other adjustment factors. Estimating  $V$  is usually possible since  $X^T(y - \pi)$  is the sum of random observations and most sample designs admit a consistent estimator of the sum of (not necessarily independent) observations. To estimate  $V$  we use  $\hat{\pi}$  instead of  $\pi$  in the estimate. Since asymptotically

$$(\hat{\theta} - \theta) \doteq (X^T \Delta X)^{-1} X^T(y - \pi),$$

we have that

$$n^{1/2}(\hat{\theta} - \theta) \rightarrow N[0, n^2(X^T \Delta X)^{-1} V(X^T \Delta X)^T];$$

see Binder (1983) for a detailed justification of this result. Now, a Wald statistic may be constructed from the estimated covariance matrix for  $\theta_2$ .

**Table 3**  
Logistic Regression Model for Explaining Use of Physician Services

Variable	Type	d.f.	Wald Statistic
Age .....	Categorical	4	19.232
Sex .....	Categorical	1	12.494
Age-Sex Interactions .....	Categorical	4	36.001
Family Income .....	Categorical	5	14.642
Occupation .....	Categorical	3	8.614
Occupation-Sex Interactions .....	Categorical	3	11.501
Marital Status .....	Categorical	3	45.752
Medical History .....	Categorical	2	36.700
Number of Health Problems .....	Quantitative	1	81.554
Drug Use .....	Categorical	2	272.175
Number of Accidents .....	Quantitative	2	106.372
Number of Disability Days .....	Quantitative	2	29.052
Community Size .....	Categorical	2	11.751
Provincial Physician - Population Ratios .....	Quantitative	1	0.540



### Example 4

A logistic regression model was fit on 20,726 respondents from the Canada health survey to explain use or non-use of physician services over a 12-month period. In total it was estimated that 77% of the population visited a physician at least once. The results are summarized in Table 3. For more complete details, see Binder (1983). The logistic model seemed to fit the data very well.

### 5.3 Qualitative Explanatory Variables

The theory of this section was obtained by G. Roberts in an unpublished manuscript (Carleton University). Here the explanatory variables are all qualitative. We label the domains,  $\{1, \dots, I\}$ . We let  $p_i$  be the survey estimate of the  $i$ -th domain proportion and  $\hat{N}_i$  is the estimate of the size of the  $i$ -th domain,  $N_i$ . Under the model, the expected proportion in the  $i$ -th domain is  $f_i$ , where

$$\log \{f_i / (1 - f_i)\} = \underline{a}_i^T \underline{\theta},$$

for  $\underline{a}_i$  known and  $\underline{\theta}$  an unknown parameter. We define  $\underline{A} = [\underline{a}_1, \dots, \underline{a}_I]^T$  and let  $\underline{D}_N = \text{diag}\{\hat{N}_1, \dots, \hat{N}_I\}$ .

Under the model, the survey estimator of  $\underline{f} = (f_1, \dots, f_I)^T$  is given by  $\hat{\underline{f}}$ , the solution to

$$\underline{A}^T \underline{D}_N (\underline{p} - \hat{\underline{f}}) = \underline{0}. \quad (5.1)$$

Since asymptotically

$$\hat{\underline{\theta}} - \underline{\theta} \doteq (\underline{A}^T \underline{\Delta} \underline{A})^{-1} \underline{A}^T \underline{D}_N (\underline{p} - \hat{\underline{f}}),$$

where  $\underline{\Delta} = \text{diag}\{N_1 f_1 (1 - f_1), \dots, N_I f_I (1 - f_I)\}$ , we have

$$n^{1/2} (\hat{\underline{\theta}} - \underline{\theta}) \rightarrow N[0, (\underline{A}^T \underline{\Delta} \underline{A})^{-1} \underline{A}^T \underline{D}_N \underline{V}_p \underline{D}_N \underline{A} (\underline{A}^T \underline{\Delta} \underline{A})^{-1}]$$

whenever  $n^{1/2} (\underline{p} - \underline{f}) \rightarrow N(0, \underline{V}_p)$ .

Under independent binomial sampling, the covariance matrix reduces to  $(N/n)(\underline{A}^T \underline{\Delta} \underline{A})^{-1}$  where  $n$  is the sample size.

The likelihood ratio test for testing goodness of fit is

$$LR_s = 2(n/\hat{N}) \sum_{i=1}^I \hat{N}_i [p_i \log(p_i/\hat{f}_i) + (1 - p_i) \log\{(1 - p_i)/(1 - \hat{f}_i)\}],$$

where  $n$  is the sample size and  $\hat{N} = \sum \hat{N}_i$ . Under  $H_0$  this is asymptotically equivalent to

$$W_{11} = (n/\hat{N}) \sum_{i=1}^I \hat{N}_i (p_i - \hat{f}_i)^2 / [f_i (1 - f_i)].$$

In general, the distribution of  $LR_s$  will be that of  $\sum \delta_i Z_i^2$ , where  $\{Z_i\}$  are independent  $\chi^2_1$  and  $\{\delta_i\}$  are the eigenvalues of  $\hat{N}^{-1} \underline{D}_N [\underline{\Delta}^{-1} - \underline{A} (\underline{A}^T \underline{\Delta} \underline{A})^{-1} \underline{A}^T] \underline{D}_N \underline{V}_p \underline{D}_N [\underline{\Delta}^{-1} - \underline{A} (\underline{A}^T \underline{\Delta} \underline{A})^{-1} \underline{A}^T] \underline{\Delta} \underline{D}_N^{-1}$ . By taking the expectation of  $W_{11}$ , and approximating

$$W_{11} \approx \frac{\sum \delta_i}{I - s} \chi^2_{I-s}$$

where  $s = \text{rank}(\underline{A})$ , we obtain

$$\sum \delta_i = (n/\hat{N}) \sum_{i=1}^I \hat{N}_i v_{ii}^{(r)} / \{f_i (1 - f_i)\}$$

where  $v_u^{(r)} = V \{ p_i - \hat{f}_i \}$ . The  $\{v_u^{(r)}\}$  may be computed using the relationship  $p - \hat{f} \doteq [I - \text{diag} \{f_i(1-f_i)\} A(A^T \Delta A)^{-1} A^T D_N] (p - \hat{f})$ .

### Example 5

The data from the October 1980 Canadian Labour Force Survey was used to fit logistic (logit) models for the probability of being employed. The sample consisted of males aged 15-64 who were in the labour force and not full time students. A logit model, quadratic in age and in education, was fitted. Age-group levels were formed by dividing the interval [15, 64] into ten groups with the  $j$ th age-group being the interval  $[10 + 5j, 14 + 5j]$ ,  $j = 1, 2, \dots, 10$ . The midpoint of each age-group was used as the value of the age for all persons in that age-group. Six levels of education were formed by assigning to each person a value based on the median years of schooling. Age by education classification led to the formation of 60 cells.

Let  $\pi_i = \text{Pr}\{\text{an individual in the } i\text{th cell is employed}\}$ ,  $i = 1, 2, \dots, 60$ . We assume that  $0 < \pi_i < 1$ . Hence  $1 - \pi_i$  represents the probability that the individual in the  $i$ th cell is unemployed. The model, considered for fit, was

$$\ln \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 a_i + \beta_2 d_i^2 + \beta_3 d_i + \beta_4 d_i^2, \quad (1)$$

$$i = 1, 2, \dots, 60$$

where  $a_i$  and  $d_i$  are the age and education variable values for the individuals in the  $i$ th cell.

Using the survey estimates  $p_i$  of  $\pi_i$ , the values of Pearson's statistic  $W_{11}$  and the likelihood ratio statistic  $LR_5$  were computed as  $W_{11} = 98.94$  and  $LR_5 = 101.20$ . The upper 5% point of the chi-square distribution, with 55 degrees of freedom, is 73.31. Using these values of  $W_{11}$  or  $LR_5$  we would reject the model. These values of  $W_{11}$  or  $LR_5$ , however, are appropriate only if the sample was a random sample.

The estimate average eigenvalue,  $\Sigma \delta/55$ , for testing goodness of fit for this data is 1.88. This would reduce  $W_{11}$  to 52.63 and  $LR_5$  to 53.83. Hence, with this adjustment, we find that the data are consistent with the model (1).

The use of the Wald statistic,  $(p - \hat{f})^T [\hat{V}^{(r)}]^{-1} (p - \hat{f})$ , for testing the goodness of fit was also considered. Here we use the  $g$ -inverse of  $\hat{V}^{(r)}$  since the matrix is singular. Some perturbation to the estimates of  $p_i$ , when  $p_i = 1$ , was necessary for computing the Wald statistic. It was found that the Wald statistic was unstable for our problem. Minor perturbations in the estimates of  $p$  led to considerable change in the value of the Wald statistic. Also the value of the Wald statistic is very large here due to instability in the estimated covariance matrix involved in its calculation. The Wald statistic is at least 30 times larger than our adjusted Chi-squared values.

## 6. SOFTWARE CONSIDERATIONS

Advancement of computer technology has made data collection, storage and retrieval operations easy and efficient. Powerful generalized software systems, such as TPL, STATPAK and ESTIMATION SYSTEM, have been used to produce cell estimates and some of their variances fairly easily to users and analysts. As well a number of commercially available packages such as BMDP, SPSS and SAS are powerful analytic tools in certain contexts. However, the ability to perform analysis such as those described in this paper are limited. For example, in situations involving hypothesis testing or statistical inference, these packages assume that the data to be analyzed come from surveys with simple random samples.

At present, an integrated software package, similar to the ones mentioned above, but designed for analyses of the type of data discussed in this paper, is not available. As a result, the researcher requiring a quick solution to his problem is usually forced to use existing statistical packages which may not be appropriate.

The alternatives are

- use existing packages with modifications
- use existing stand-alone software
- write customized programs
- use combinations of the above.

For the analyses given in this paper, modifications to the MINI CARP program (Hidioglou, Fuller and Hickman; 1980) were incorporated to obtain the results in Examples, 1, 2 and 3. For Example 4, a combination of PL/1 and SAS programs were developed. The analysis of the Labour Force Survey data (Example 5) used a combination of customized programs and SAS.

For the above alternatives, some practical drawbacks have been experienced. they include:

- (a) If an existing package is to be modified, intimate knowledge of the package is often required;
- (b) Identical information may have to be duplicated on separate data files, as these alternatives are not integrable like generalized systems;
- (c) Compared to an integrated "user-friendly" package, these alternatives lack elegance and operational efficiency as software;
- (d) Comprehensive documentation is not generally available for specially written programs limiting the availability of software.

Work is now ongoing to develop SAS based procedures for performing many of these analyses. Our ultimate goal is similar to that proposed by Shah (1981); namely, the development of an integrated software package for survey data analysis. This is a goal worth striving for, if we are to avoid the frustrations now being experienced by researchers who are faced with either developing their own software or using existing software which could lead to erroneous results and conclusions.

## 7. DISCUSSION

We have examined a number of problems which arise when fitting models to categorical data which have been collected under complex sampling designs. The basic approach has been to derive the appropriate Wald statistic for the fitted model or to use the test statistic which is motivated from multinomial-type sampling designs and find a suitable approximation to its null distribution.

We have not addressed the issue as to whether one should really be taking a model-based or design-based approach to begin with. Instead, we have concentrated on design-based inferences.

To put this issue into focus, let us reconsider the test of independence in a two-way contingency table. The question of independence arises if we are interested in whether knowing the value of variable  $Y_1$  affects our knowledge about variable  $Y_2$ . If it does not, for all the individuals in the population, then we say the variables are independent. However, if we also know the value of  $Y_3$ , it may turn out that  $Y_1$  and  $Y_2$  are no longer independent. This is particularly important when  $Y_3$  is a design variable (such as geographic stratum). Since design variables are usually known for all sampled individuals, we have one of two options: (a) we can say that the question of independence is no longer relevant, or (b) we can marginalize out  $Y_3$ , and say that we are only interested in  $Y_1$  and  $Y_2$ , unconditionally. Assuming that we take approach (b), the results of this paper seem appropriate. In some cases it may be possible to test if  $Y_1$  and  $Y_2$  are conditionally independent given  $Y_3$ .

There is a further difficulty, however. Suppose we are interested in the cell proportions  $\pi_{ij}$  from a finite population of size  $N$ . If we were to take a census from this population, it is highly unlikely that we would obtain  $\pi_{ij} = \pi_{i+} \pi_{+j}$  exactly. The best that we could hope for is that some measure of association such as  $N \sum (\pi_{ij} - \pi_{i+} \pi_{+j})^2 / \pi_{i+} \pi_{+j}$  is small. Note that ever

under a super-population model of exact independence, we would not expect this measure of association to be zero. Perhaps, we should instead be testing hypotheses such as

$$H_0: \text{Measure of Association} \leq C$$

$$H_1: \text{Measure of Association} > C.$$

Further research is needed in this area. However, for practical circumstances where the sampling fraction is not large, the methods given in this paper are suitable.

## REFERENCES

- ALTHAM, P.A.E. (1976). Discrete variable analysis for individuals grouped into families. *Biometrika*, 63, 263-269.
- BINDER D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: MIT Press.
- BRIER, S.S. (1978). Discrete Data Models with Random Effects. Technical Report, University of Minnesota, School of Statistics.
- COHEN, J.E. (1976). The distribution of the chi-squared statistic under cluster sampling from contingency tables. *Journal of the American Statistical Association*, 71, 665-670.
- COWAN, J. and BINDER, D.A. (1978). The effect of a two-stage sample design on tests of independence in a 2 by 2 table. *Survey Methodology*, 4, 16-28.
- DAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of American Statistical Association*, 80, 148-157.
- ELLEG, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of American Statistical Association*, 71, 665-670.
- FIENBERG, S.E. (1980). *The Analysis of Cross Classified Data*, (2nd ed.). Cambridge, Massachusetts: MIT Press.
- RIZEL, J.E., STARMER, C.F. and KOCH, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, 25, 489-504.
- IDIROGLOU, M.A., FULLER, W.A. and HICKMAN, R.D. (1980). MINICARP: A program for estimating simple descriptive statistics and their variances for multi-stage stratified designs. Iowa State University: Ames, Iowa.
- IDIROGLOU, M.A. and RAO, J.N.K. (1981). Chisquare tests for the analysis of categorical data from the Canada Health Survey. Paper presented at the International Statistical Institute Meetings, Buenos Aires, 1981.
- IDIROGLOU, M.A. and RAO, J.N.K. (1983). Chi-square tests for the analysis of three-way contingency tables from the Canada Health Survey. Technical Report, Statistics Canada.
- IRY, P.B., KOCH, G.G. and STOKES (1981). Categorical data analysis: Some reflections on the log linear model and logistic regression. part I: Historical and Methodological Overview. *International Statistical Review*, 49, 265-283.
- JOHNSON, N.L. and KOTZ, S. (1970). *Continuous Univariate Distributions*. Boston: Houghton Mifflin.
- KOCH, G.G., FREEMAN, D.H. JR., and FREEMAN, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. *International Statistical Review*, 43, 59-78.
- RAO, J.N.K. and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-square tests for goodness of fit and independence in two-way tables. *Journal of American Statistical Association*, 76, 221-30.
- RAO, J.N.K. and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 48-60.

- SATTERTHWAITE, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.
- SHAH, B.V. (1981). Development of survey data analysis software. Research Triangle Institute, Research Triangle Park, North Carolina.
- SHUSTER, J.J. and DOWNING, D.J. (1976). Two-way contingency tables for complex sampling schemes. *Biometrika*, 63, 271-276.



## Estimating Economic Cycles in Semi-Annual Series

PIERRE A. CHOLETTE<sup>1</sup>

### ABSTRACT

This paper presents a moving average which estimates the trend-cycle while eliminating seasonality from semi-annual series (observed twice yearly). The proposed average retains the power of all cycles which last three years or more; 90% of those of two years; and 55% of cycles of one year and a half. By comparison, the *two by two* moving average retains the power of respectively 75%, 50% and 25% of the same cycles.

KEY WORDS: Moving averages; Economic cycles; Spectral analysis.

### 1. INTRODUCTION

In some cases, semi-annual series exist for which there are no corresponding monthly data. In such instances, one cannot derive the seasonally adjusted semi-annual series from the monthly seasonally adjusted values. In addition, to our knowledge, there are no seasonal adjustment methods for semi-annual series.

This paper presents a moving average which eliminates seasonality and estimates the trend-cycle of semi-annual series. The approach of quadratic minimization used originates with Whittaker (1923) and was further developed by Leser (1961 and 1963), Cholette (1980), Schlicht (1981) and others.

The average derived has five terms and comprises a set of central weights for the semestres half-years) at the centre of series; and two sets of end weights, for the two first and last semestres. Consequently there is no loss of estimates at the ends of series, as with the *two by four* moving average (used for quarterly series) for instance.

The spectral properties of the central weights prove to be superior to those of the *two by two* moving average, which first comes to mind as a way of processing semi-annual series. The properties of the end weights will also be examined.

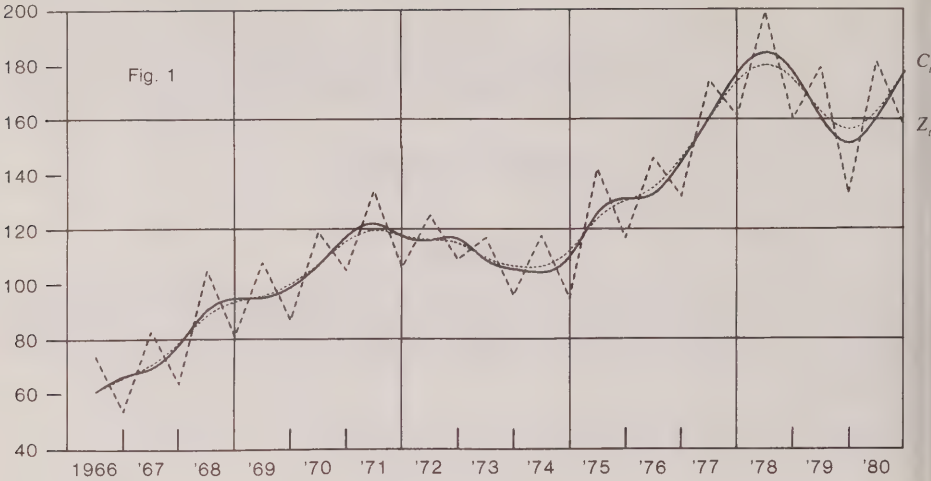
### 2. ILLUSTRATION OF THE AVERAGE

Figure 1 shows the observed semi-annual original series  $z_t$  (dashed line) along with the trend-cycle  $c_t$  (solid line) estimated by the semi-annual cyclical average presented in this paper. As expected, the trend-cycle behaves smoothly and displays short run cycles, namely a three-year cycle extending from the second semestre (half-year) of 1977 to the first of 1980. An estimate is available for each observation, including the two first and last observations. The trend-cycle produced by the *two by two* moving average (dotted curve) on the other hand does not yield any estimate for the first and the last semestres. Furthermore, the *two by two* does not reach as deeply into the cyclical peaks and troughs compared to the proposed average.

<sup>1</sup>Pierre A. Cholette, Time Series Research and Analysis, Statistics Canada, 25th floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.



$Z_t, C_t$



**Figure 1.** Semi-annual seasonal series (----) and its trend cycle (—) estimated by the proposed semi-annual cyclical moving average and by the *two by two* moving average (····)

3. WEIGHTS OF THE AVERAGE

Table 1-A displays the exact values of the weights of the semi-annual cyclical moving average used. The first row gives the *modified* central weights pertaining to estimates 3 to 28 (in Figure 1); the second, the end weights pertaining to the second-last estimate; and, the third, to the last estimate. Table 1-B shows the central weights derived according to the methodology of Section 6. We judged however these should be replaced by the *modified* central weights of Table 1-A for reasons to be explained.

**Table 1-A**  
Exact weights of the proposed semi-annual cyclical average

Modified					
Central weights	-0.1000	0.2500	0.7000	0.2500	-0.1000
Second-last					
Set of weights	0.0625	-0.2500	0.3750	0.7500	0.0625
Last					
Set of weights	-0.0625	0.2500	-0.3750	0.2500	0.9375

**Table 1-B**  
Unmodified central weights

	-0.0625	0.2500	0.6250	0.2500	-0.0625
--	---------	--------	--------	--------	---------

#### 4. SPECTRAL ANALYSIS OF THE AVERAGE

The curves of Figures 2 and 3 represent the gain functions of the set of weights analysed. The gain value on the ordinate indicates the percentage of amplitude of the sinusoidal waves preserved, that is *passed* to the estimates, by the weights. The frequency of these waves is shown on the abscissa and varies from 0 to 0.500. Frequency 0.500 corresponds to the annual wave of two semestres (1/.50), that is to stable seasonality. Moving seasonality is accounted for by a few neighbouring quasi annual frequencies: 0.467 and 0.483.

Frequency 0.333 corresponds to a three (1/.33) semestre wave; frequency 0.250, four semestres; 0.200, to five semestres; 0.167, six semestre, etc. Frequencies associated with waves of three semestres or more (left of 0.333 in figures 2 and 3) pertain to the trend-cycle of series and constitute the target frequencies of the estimator.

The frequencies between 0.333 and 0.467 exclusively are associated with fluctuations of periodicity less than one and a half years and superior to the quasi-annual seasonal frequencies. They pertain to the irregular component of series. An ideal cyclical average should eliminate 100% of these irregular frequencies, 100% of the seasonal and quasi-seasonal frequencies and preserve only the cyclical frequencies from 0 to 0.333 inclusively.

##### a) Analysis of the central weights

The solid curve of Figure 2 shows that the modified central weights of the semi-annual cyclical average preserves 100% of all waves of five semestres (2 years) and more: everywhere left of frequency 0.200 the curve is above 100%. By comparison, the *two by two* moving average, represented by the dotted curve, only preserves 65% of the 5-semestre waves and 93% of the 10-semestres waves. Furthermore, the modified central weights pass 55% of 3-semestre waves and 90% of 4-semestre (2-year) waves; against 25% and 50% respectively for the two by two.

Both sets of weights completely eliminate stable seasonality, with gain valued at zero for the seasonal frequency 0.500; and nearly all the moving seasonality. However, the two by two eliminates slightly more of the irregular frequencies than the modified central weights. When choosing between the two averages, one then faces the following trade-off: to let the estimates contain more cyclical movements but also more irregularity or less cyclical movements and less irregularity. When a series is known to contain more cyclical movements (especially faster movements) than irregularity, the modified central weights of the proposed average are certainly preferable to the two by two.

The dashed curve of Figure 2 represents the gain of the unmodified central weights of the semi-annual cyclical average, as obtained from Section 6. At the cyclical frequencies, its performance proves superior to that of the *two by two*; but, inferior to that of the modified central weights. For instance, the latter reproduces 101% of the 5-semestre waves (frequency 0.200) against 88% for the unmodified. The amplification of 5% (gain of 105%), at the 6-semestre frequency 0.167) wave with the modified central weights, seems to us preferable to a comparable reduction of 6% (gain of 94%) with the unmodified set of weights. Indeed the analyst stands a better chance to detect an amplified signal in a series than a reduced signal.

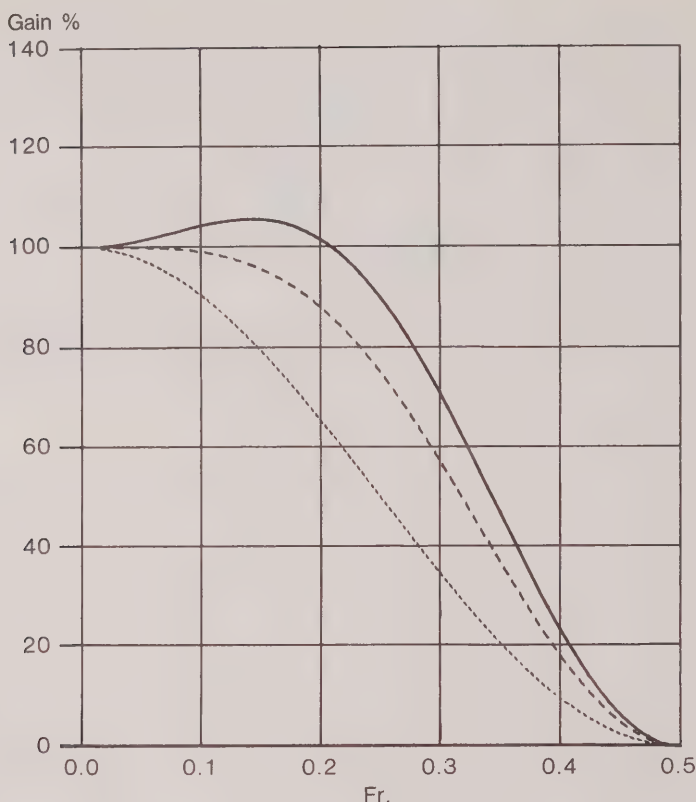
##### b) Analysis of the end weights

Ideally, the gains of the end weights should be identical to the gain of the central weights. In such a case, end and central weights would have the same effect on the processed series (except for possible phase-shifts).

The gain of the weights for the second-last estimate (dotted line of Figure 3) is quite similar to the gain of the modified central weights (solid curve). Note that the former is more similar to the latter than to the set of unmodified central weights of Figure 2. This is the reason why we modified the weights.

The weights for the second-last estimate preserve the cyclical frequencies and eliminate stable seasonality. However, they preserve 11 and 21% of the moving seasonality frequencies 0.467 and 0.483; and, even more of the noise frequencies. From the view point of the gain, the weights of the second-last estimate should yield less reliable estimates than the modified central weights.

Fig. 2



**Figure 2.** Gain functions of the modified (—) and non-modified (----) central weights of the proposed semi-annual cyclical average and of the *two by two* average (····)

The situation gets worse for the set of weights for the last estimate (broken line in Figure 3). Here, a strong amplification of some noise and fast cyclical frequencies is observed (gains reaching up to 137%). Caution should then be exercised in interpreting the estimate yielded by these weights. One should perhaps disregard the last estimate completely, for series which are reputed to be irregular (containing those magnified frequencies).

As seen in Table 1, the end sets of weights are not symmetric. Consequently, they cause phase-shifts, which are compiled in number of semestres in Table 2 for certain selected frequencies. At the target cyclical frequencies, a small phase-shift is observed for the second-last weights. In this case, a cyclical wave of five semestres will be delayed by 0.09 semestres in the estimates; one of four semestres, by 0.16 semestres; and one of three semestres, by 0.28 semestres; etc.

The phase-shift reaches its maximum at the fundamental seasonal frequency (0.500). This does not matter however, since the frequency is totally eliminated by the weights. It does matter a little for the moving seasonality frequencies 0.467 and 0.483, since they are not completely eliminated.

Fig. 3

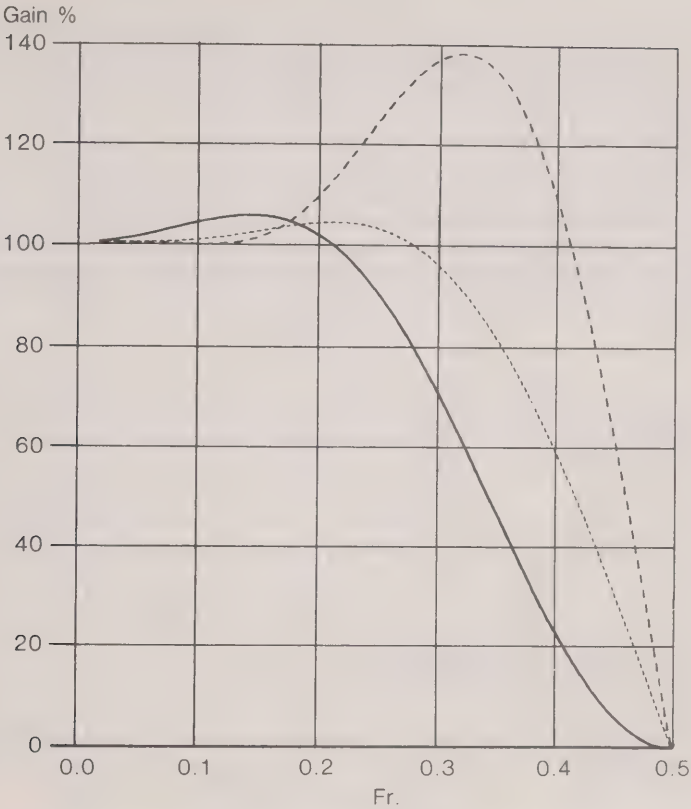


Figure 3. Gain functions of the modified central weights ( — ) and of the weights for the second-last ( ··· ) and the last ( --- ) estimates

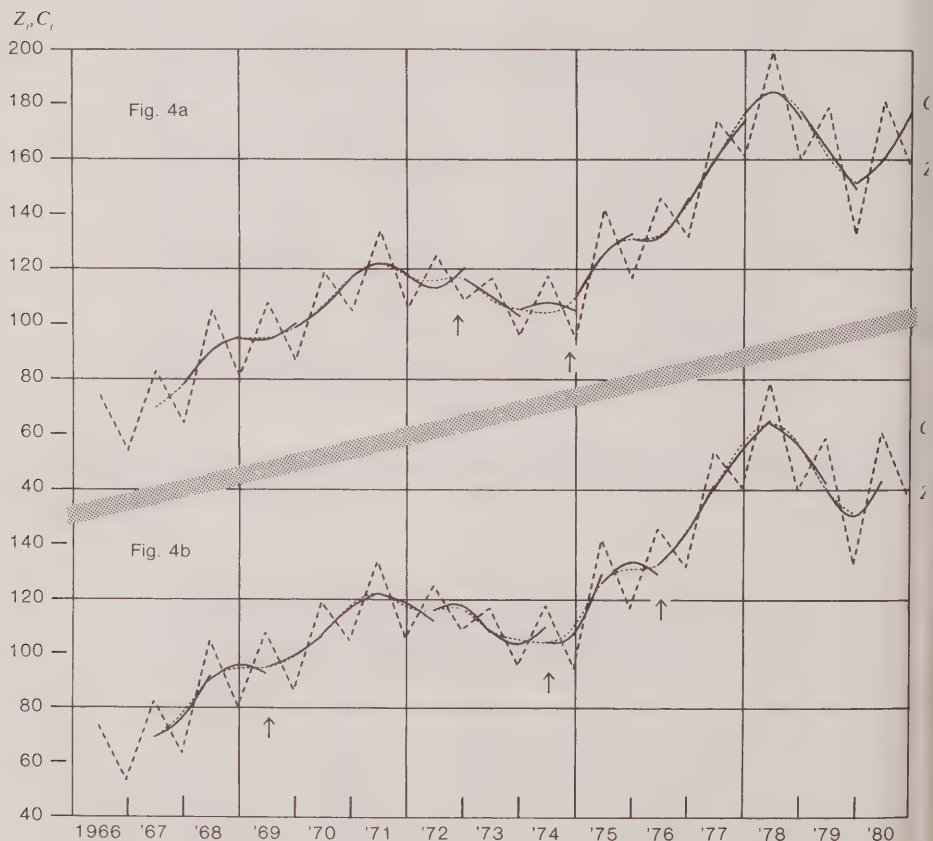
Table 2  
Phase-shifts observed in number of semestres for the sets of  
end weights at certain selected frequencies

	second-last	last
cyclical frequencies:		
0.100 ( 10 semestres)	0.01	0.01
0.167 ( 6 semestres)	0.05	0.05
0.200 ( 5 semestres)	0.09	0.05
0.250 ( 4 semestres)	0.16	0.00
0.333 ( 3 semestres)	0.28	0.17
seasonal frequencies		
0.467	0.46	0.45
0.483	0.48	0.47
0.500 ( 2 semestres)	0.50	0.50

### 5. GRAPHICAL ANALYSIS OF END ESTIMATES

Figure 4 displays the preliminary estimates derived using the two sets of end weights for years 1968 to 1980, accompanied by the corresponding central final available estimates. Figure 4 a) shows the end estimates falling in the second semestre; and 4 a), in the first semestre. (One single plot would have been too crowded.)

If the central estimates are considered as true (or at least more reliable, the end estimates are seen to cause five false signals: in 1968 (arrow in fig. 4 b), in 1972 (4 a), in 1974 (4 b), in 1975 (4 a) and in 1976 (4 b). A false signal is said to occur here when the end estimates show a change in the direction of the trend-cycle and when that change is later contradicted by the central final estimates (becoming available with new observations). These false signals tend to appear when the series slows down in one direction and resumes the movement in the same direction. When there is a strong change of direction like in 1978, this does not seem to occur.



**Figure 4.** Semi-annual seasonal series (----); preliminary estimates of its trend-cycle by the ends weights (—) of the proposed semi-annual cyclical moving average a) for the second semestres and b) for the first semestres; final estimates (····) by the central weights of the average.



If the estimates derived by the last set of weights were omitted, many false signals would *disappear*. However, the estimates could become less timely. This illustrates the statistician's dilemma between the timeliness and the reliability of estimates under any estimation method. (In practise, a serious analyst would wait for at least one confirmation of a signal before *believing* it).

Apart from the five false signals mentioned, the preliminary estimates display a movement which is very similar and sometimes undistinguishable from that of the final estimates.

5. CALCULATION OF THE WEIGHTS OF THE SEMI-ANNUAL CYCLICAL AVERAGE

The observed series  $z_t$  comprises the trend-cycle  $c_t$  to be estimated and a seasonal-irregular residual  $s_t + e_t$  ( $= z_t - c_t$ ):

$$z_t = c_t + (s_t + e_t), t = 1, \dots, 5. \tag{1}$$

Following the approach of Leser (1961 and 1963) and of Cholette (1980), the desired trend-cycle minimizes the quadratic sum of fourth differences (first term of (2)). On the five-semester estimation interval, the component as much as possible approximates a time polynomial of the third degree. This specification allows for a full economic cycle with its four phases of expansion, turning-point, recession and recovery over the interval.

The seasonal-irregular residual ( $z_t - c_t$ ) minimizes the quadratic sum of first seasonal differences taken on corresponding semestres (second term of (2)). This specification means that the seasonal-irregular residual of one semestre should resemble that of the same semestre in the neighbouring year as much as possible.

Furthermore, the seasonal-irregular residuals minimize the quadratic sum of their sums on two consecutive semestres (third term of (2)). This criterion indicates that the seasonality of two neighbouring semestres should cancel out and that the irregularity should not affect the level of the desired trend-cycle.

The three criteria specified for the components combine into the following objective function:

$$f(c) = \sum_{t=5}^5 (c_t - 4c_{t-1} + 6c_{t-2} - 4c_{t-3} + c_{t-4})^2$$

$$+ \sum_{t=3}^5 \{(z_t - c_t) - (z_{t-2})\}^2 + \sum_{t=2}^5 \{(z_t - c_t) + (z_{t-1} - c_{t-1})\}^2$$

Equation (3) can be rewritten in linear algebra:

$$f(C) = C' A' A C + (Z - C)' B' B (Z - C) + (Z - C)' F' F (Z - C)$$

$$= C' H C + (C - Z)' G (C - Z)$$

where A, B and C respectively stand for the matrix operators of quadruple differences, first seasonal differences and annual sums defined as follows:

$$A = \begin{bmatrix} 1 & -4 & 6 & -4 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix}, \quad F = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The normal equations associated with (3) read

$$dF/dC = 2HC + 2G(C - Z) = 0 \tag{4}$$

and imply solution:

$$C = (H + G)^{-1} G Z = W Z \tag{5}$$



The third central row of matrix  $W$  contain the non-modified central weights of the semi-annual cyclical average of Table 1-B; and the fourth and fifth row, the end weights of rows 2 and 3 of Table 1-A.

## 7. HISTORY OF MOVING AVERAGES BY QUADRATIC MINIMIZATION

This approach of quadratic minimization originates with Whittaker (1923). Leser (1961, 63) showed how quadratic minimization could be applied to develop cyclical moving averages. Cholette (1980) proposed substitutes for the two by twelve and the two by four moving averages. These substitutes were incorporated into the Dagum (1980) seasonal adjustment programme as optional.

The semi-annual cyclical average presented in this paper could also be incorporated in a seasonal adjustment method of the X-11 type. This would allow seasonally adjusting semi-annual series and the calculation of the seasonal factors by means of the seasonal moving averages usually applied for monthly and quarterly series.

## 8. CONCLUSION

This paper presented a 5-term moving average which eliminates seasonality from semi-annual time series. The estimator reproduces the economic cycles more exactly than the two by two moving average. The two by two also has the disadvantage of not providing any estimate for the first and last semestres of the series.

## REFERENCES

- AKAIKE, H. and ISHIGURO, M. (1980). BAYSEA, a Bayesian Seasonal Adjustment Program. *The Institute of Statistical Mathematics*, Computer Science Monograph No. 13, Tokyo.
- BOX, J.E.P., and JENKINS, G.M. (1970). *Time Series Analysis, Forecasting and Control* Holden-Day.
- CHOLETTE, P.A. (1980). A Comparison of Various Trend-Cycle Estimators. in *Time Series Analysis*, (O.D. Anderson and M.R. Perryman Eds.), 77-87.
- DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method*. Cat. 12-564E Statistics Canada.
- KOOPMANS, L.H. (1974). *The Spectral Analysis of Time Series*. Academic Press.
- LAROCQUE, G. (1977). Analyse d'une methode de dsaisonnalisation: le programme X-11 du U.S. Bureau of the Census, version trimestrielle. *Annales de l'I.N.S.E.E.*, 28, 105-127.
- LESER, C.E.V. (1961). A Simple Method of Trend Construction. *Journal of the Royal Statistical Society* Ser. B, 23, 91-107.
- LESER, C.E.V. (1963). Estimation of Quasi-Linear Trend and Seasonal Variation. *Journal of the American Statistical Association*, 58, 1033-1043.
- MACAULY, F.R. (1931). *The Smoothing of Time Series*. National Bureau of Economic Research. Washington.
- PHILIPS, L. BLOMME, R. (1973). *Analyse chronologique*, (Vander Eds.). Louvain, 334.
- SCHLICHT, E. (1981). A Seasonal Adjustment Principle and a Seasonal Adjustment Method Derived from this Principle. *Journal of the American Statistical Association*, 76, 374-378.
- SHISKIN, J., YOUNG, A.H. and MUSGRAVE, J.C. (1967). *The X-11 Variant of Census the Methoa II Seasonal Adjustment Program*. Technical Paper No. 15, U.S. Bureau of the Census.
- WHITAKKER, E. (1923). On a New Method of Graduation. *Proceedings of the Edinburg Mathematical Society*, 41, 63-75.

## **The Use of Matching in the Evaluation of Non-Sampling Errors in the 1981 Canadian Census of Agriculture**

**J. COULTER<sup>1</sup>**

### **ABSTRACT**

This paper discusses the use of matching between files of comparable data in the evaluation of non-sampling error. As an example of the technique, the data quality evaluation of the 1981 Canadian Census of Agriculture is described and some results presented.

**KEY WORDS:** Non-sampling error; Coverage; Response error; Matching; Record Linkage; Census of Agriculture.

### **1. INTRODUCTION**

As the use of probability sampling in data collection has evolved, the evaluation and control of sampling errors has been a constant concern. Extensive research has been devoted to the design of sampling schemes which would reduce sampling error and facilitate its measurement. In many situations, however, major portions of the survey error arise not from sampling, but from the effects of other components of the data collection operation. In censuses particularly, in which data are obtained through 100 percent enumeration of the population of interest, sampling error is nonexistent. Instead survey error is due entirely to the influences of respondents, interviewers, coders, keyers, and others during the collection, capture, and processing stages of the survey operation. As the impact of these non-sampling errors on data quality has become more fully understood, the development of techniques to control and measure them has gained in importance.

### **2. MODELS FOR SURVEY ERROR**

Early papers on total survey error, such as that by Deming (1944), outlined the potential sources of error and discussed the need to consider their varying effects when planning data collection operations. As the study of survey error developed, general models were proposed by Hansen et al. (1951), Sukhatme and Seth (1952), Hansen, Hurwitz, and Bershada (1961), and others to describe the components of sampling and non-sampling error. Studies were conducted on the correlations between errors which result from influences such as interviewers or coders, and methods were developed for measuring their effects. Fellegi (1964) presented a detailed model which included correlations between numerous error sources.

Other models have followed which consider both single and correlated non-sampling errors and propose methods for evaluating them. Some examples include the U.S. Bureau of the Census survey error model described by Nisselson and Bailer (1976), the discussion of measure-

---

<sup>1</sup> J. Coulter, Census Operations Division, Statistics Canada, 2nd Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

ment errors by Cochran (1977), and the model of survey error presented by Andersen et al. (1979) which was based on an earlier model by Kish (1965). In a recent paper Hartley (1981) described a model with terms for interviewer, coder, and respondent errors, and proposed a sample design to facilitate estimation of these errors.

Throughout the literature the components of non-sampling error have been categorized in a variety of ways. In this paper we will divide non-sampling error into two elements, coverage error and response error. A coverage error will be defined to have occurred when a unit which satisfies the definition of the universe of interest is missed or counted more than once, or a unit not belonging to the desired universe is included. Coverage errors cause the invalid inclusion or exclusion of all data for the incorrectly enumerated unit. As a result they may influence the estimates for any or all data items.

Response errors will be defined as affecting the values of individual items within the data for units which have been correctly included in the enumeration. They may have arisen at the initial collection of the data or during subsequent processing stages; potential sources include misinterpretation of a question by the respondent, total or partial non-response, the influence of the interviewer, and data capture or coding error.

### 3. EVALUATING NON-SAMPLING ERROR

In most survey error models, error is defined as the difference between the true value and the collected value for the particular data item. Thus in order to evaluate non-sampling error, one would in theory propose to compare the data collected by the survey with the true values for the item of interest. In practice, however, the true values are seldom known, if in fact they even exist. Instead the survey data must be compared to estimates from an alternate source which is believed to provide the closest available approximations to the true values.

In determining the data source most suitable to represent the unknown true values, a number of factors must be considered. The alternate data must be collected independently from those of the survey of interest. Optimally, the definitions and concepts employed in the collection of the data, and the reference periods to which the data applied, would be equivalent for the two sources. The universe covered by the alternate data would be the same as that of the survey, or comparable subuniverses would be identifiable. The purpose and methods of collection of the alternate data, and of any subsequent processing or updating stages, would be fully understood. Perhaps most important of all, the data would be of high enough quality to act as a standard against which the survey data could be compared.

In practice, of course, all of these conditions are seldom satisfied by a single alternate data source. In some cases one source may be best for a particular subpopulation while another is preferable for the remainder of the universe. Adjustments may be possible to remove the effects of differences in reference dates or definitions of variables between the two data sources. For the majority of cases, however, even the best estimator of the true values will involve major failures of some of these conditions, and their influence on the comparison may not be measurable or even identifiable.

Approximations to the true values may be obtained from a variety of sources. Estimates from one census or survey may be compared to those from another. Examples include the comparisons of the Current Population Survey and the U.S. Census of Population, the Labour Force Survey and the Canadian Census of Population, and the Agriculture Enumerative Survey and the Canadian Census of Agriculture (Statistics Canada 1979). Demographic projections have also been employed to approximate the true values, such as in the evaluations of the Labour Force Survey and the Canadian Census of Population described by Fellegi (1973).

Administrative data are also being used more and more in evaluation studies as respondent burden becomes an increasing concern. Estimates from income tax, family allowance, motor vehicle licence, agricultural marketing board, and other files may provide approximations to the required true values. Income tax files, for example, are currently being studied at Statistics Canada for use in the collection and evaluation of farm income data.



As these few examples imply, a wide range of alternate data sources have been used as standards of comparison for survey estimates. However, while macro-level comparisons provide indications of the total error in the survey results, they cannot identify or measure the components of the error. The effects of coverage and response errors cannot be identified. Hence additional methods are required which facilitate more detailed investigations into the total error observed.

One technique, which can provide the analyst with a wealth of information on errors and their sources, links the survey results at the individual record level to a file of comparable data. Using the matched records, one-to-one comparisons can be made between the values reported for the survey and those from the alternate data source, which are assumed to represent the true values. Cross-classifications over other data items provide insight into the characteristics of units displaying certain types of inconsistencies. As well, the study of records which could not be matched can indicate areas of potential over or undercoverage by the survey.

In this paper we will consider the use of matching in the evaluation of survey non-sampling error. The strengths and weaknesses of the technique, and the types of studies which it makes possible, will be discussed. As an example of the use of matching, the data quality evaluation of the 1981 Canadian Census of Agriculture will be outlined and some results presented.

#### 4. THE USE OF MATCHING

Study of the literature on the evaluation of total survey error reveals many studies which have made use of matching. As with the macro-level comparisons, a wide variety of comparable data sources have been employed. Post enumeration surveys, aimed at collecting data of a higher quality and in more detail than the original survey, have been conducted in a number of situations. Examples include reinterview by appraisers in a study on the reporting of the market value of homes (Kish and Lansing 1954), the post enumeration survey used to evaluate the U.S. Census of Agriculture (U.S. Bureau of the Census 1982), the Labour Force Survey reinterview program (Tremblay, Singh, and Clavel 1976), and the Vacancy Check Operation of the Canadian Census of Population and Housing (Statistics Canada 1980).

Independently-collected censuses and surveys have also been linked in order to evaluate data quality. Matches have been performed, for example, between the Labour Force Survey and the Canadian Census of Population (Krotki 1980), the Current Population Survey and the U.S. Census of Population (U.S. Bureau of the Census 1964), and the Agriculture Enumerative Survey and the Canadian Census of Agriculture (Statistics Canada 1979).

With the concern for reducing respondent burden, administrative data have been used increasingly in evaluation studies. Records of doctors and hospitals have been matched to health survey results by Andersen et al. (1979) and Horvitz (1981). Immigration and birth records have been employed in the Reverse Record Check Content Study of the Canadian Census of Population (Krotki 1980), and tax records of the IRS have been used to study response errors in the U.S. Census of Population (U.S. Bureau of the Census 1970). Other examples of linkage with administrative data include the evaluation of agricultural survey results by Faulkenberry and Tortora (1981) and the reporting of sensitive topics by Marquis, Marquis, and Polich (1981).

In general, the quality of an evaluation study based on a record linkage operation depends on two major factors:

- ) the quality of the data on the alternate source to which the survey is matched, and
- ) the uniqueness of the identifiers used for the match, and the accuracy of the match technique itself.

Firstly, by definition of the study objectives, the alternate data are to act as approximations to the true values. Random errors in the data which tend to cancel one another out over all records do not noticeably affect macro-level comparisons. However such errors can have a serious effect on studies conducted at the individual record level.

Some data bases chosen to act as standards for comparison may be assumed to be free from error. Birth records, for example, provide very accurate data on place of birth and age. Other data sets may be known to contain certain response or coverage errors, but if the errors are measurable they can be taken into account in the analysis, and the assumption of no error remains valid. In many cases, however, the comparable data are subject to errors which cannot

be completely identified or measured. In these situations, the best approximation to the true values is provided by the data set which is least affected by error. Data which have been collected using more accurate methods or better trained staff than the survey of interest may be assumed to contain less error, and hence can provide a reasonable basis for comparison.

The second major factor affecting the quality of the study is a function of the match operation itself. In some situations, each member of the population will have been assigned a unique identification number which has been accurately stored on both files. Linking the records would then be a straightforward process of matching on these unique identifiers. At the other end of the scale, the best available identifiers may be non-unique characteristics, such as name, which are prone to the introduction of error during data collection or capture.

The linkage algorithm itself can also have an impact on the quality of the match, particularly when the keys or identifiers are less than perfect. The algorithm may tend to allow invalid matches between records with similar keys, or may prevent valid matches when the keys differ due to minor errors or omissions. The extent to which such errors occur can have a significant effect on the composition of the files of matched and unmatched records.

Other factors which affect the comparability of the two sets of data, as for macro-level studies, include differences in collection date and method, concepts and definitions, and reference period. Due to the greater detail of investigation and cross-classification over related variables which is involved in micro-levels studies, such differences can have a much greater impact on the analysis than for the macro-level comparisons.

In order to study the use of matching in the analysis of non-sampling error, we will now consider the example of the data quality evaluation of the 1981 Canadian Census of Agriculture. For this study, independently-collected agricultural data for macro and micro-level comparisons were provided by the Agriculture Enumerative Survey (AES) and the Farm Enumerative Survey (FES), annual probability surveys conducted by Statistics Canada.

## 5. COMPARING THE CENSUS AND SURVEYS

The Canadian Census of Agriculture was conducted on June 3, 1981, sharing field operations with the quinquennial Census of Population and Housing. Data were to be collected for every census farm in Canada, defined as any farm, ranch or other agricultural operation which received \$250 or more from the sale of agricultural products during the twelve months prior to census day, or which had the potential to produce that value in the next twelve months. During drop-off of the population and housing questionnaire, the census representative was to ask at each household whether any member operated a farm or other holding which satisfied the above definition. If so, a Census of Agriculture form was left to be completed by the operator.

In order to improve coverage, results of the 1976 Census and subsequent agricultural surveys were used to identify farms which were major producers of one or more specified agricultural commodities. The census representatives then had to account for each of these "specified farms" located in the area to be enumerated.

The questionnaire, delivered prior to June 3 to the operator of each census farm, was to be completed by self-enumeration on census day. Items covered in the census included crops, livestock, land use, sales, expenses, and other areas of interest to the public and private sectors. (Further details on the methodology and content of the 1981 Census of Agriculture may be obtained from the publication Statistics Canada (1982).)

The Agriculture Enumerative Survey (AES) and Farm Enumerative Survey (FES) together covered the majority of Canada's agricultural land. The FES enumerated the Prairie provinces of Manitoba, Saskatchewan, and Alberta plus the Peace River district of British Columbia, and the AES covered the remainder of British Columbia and the provinces of Prince Edward Island, Nova Scotia, New Brunswick, Quebec, and Ontario. The survey universe consisted of agricultural holdings which satisfied the census farm definition described above. However it excluded types of organization which were of marginal economic influence, such as institutional

farms, and areas which contain little or no agricultural activity, such as urban cores. In order to provide comparable universes for the evaluation, the census file was adjusted by removing operations of these types. The deletions consisted of only 2.8 percent of the farms and 1.8 percent of the total farm area from the complete census file.

The probability surveys collected data on the same major agricultural variables as the census, such as crops, livestock, land use, and operating expenses, and used similar concepts and definitions. Some differences existed in wording and format, and in the instructions on what to include or exclude, for particular questions. As will be indicated in the discussion of the results, the effect of these inconsistencies had to be taken into consideration when comparing data from the two sources.

The AES and FES were conducted on July 1, 1981, approximately one month after the June 3 census date. Some data, such as farm expenses for the previous year, were expected to be relatively unaffected by the difference in reference data. However, other items were more likely to change between June 3 and July 1. The effect was expected to be particularly significant for livestock items, due to the constant fluctuations in inventories caused by birth, deaths, purchases, transfers, etc. As a result, operators responding to the survey were asked to indicate the changes in numbers of cattle and pigs between June 3 and July 1. Evaluation indicated that, while the data obtained were of some use in reconciling the differences due to reference date, they were subject to high non-response and questionable accuracy. Hence, the comparison of these and any other variables which would tend to be influenced by date of response had to take into consideration the difference in reference dates between the census and survey.

The samples for the AES and FES were selected from an area frame of agricultural enumeration areas, supplemented by a list frame of farms which were major producers of certain important commodities. Data collection was performed by trained enumerators during a personal interview with the operator of each selected farm. Following the necessary processing stages, an estimation procedure was applied to scale the counts up to the level of estimates for the entire population of interest. (Further details on the sample design are found in Statistics Canada (1984) and Phillips (1978).)

The survey estimates were subject to the same types of non-sampling errors as those from the census. However, due to the concentration on a smaller number of holdings, and the improved control of operations which was thus possible, it was expected that these types of errors would have a lesser impact on the surveys. Hence the surveys provided acceptable approximations to the true data values. On the other hand, the survey estimates were affected by sampling error, which had to be taken into account when making comparisons with macro-level estimates obtained from the census.

### 5.1 Macro-level Comparisons

Prior to the evaluation using the matched file, estimates from the complete census and survey data files were studied. Since the two vehicles covered comparable universes, these macro-level comparisons for provinces and regions provided initial indications of census coverage. By comparing census point estimates with survey 95 percent confidence intervals for totals of livestock, crop acreages, and other items, areas of potential over or underestimation were identified. Further investigation of the macro-level differences was then initiated to determine if they were confined to particular categories of the items of interest. The macro-level studies, in addition, provided the experience and familiarity with the two sets of data which were required for the detailed analysis which followed.

As an example of the results of the macro-level comparisons, Table 1 presents estimates for Canada for the number of farms, total farm area, and land use. A significant difference between census and survey estimates was observed for total farm area in Canada, yet the size and direction of differences varied greatly among the component land use categories. Census estimates for classes of improved land differed from the survey estimate by as much as  $25.5 \pm 7.7$  percent for other improved land to as little as  $-3.4 \pm 2.2$  percent for cropland. The



**Table 1**  
Comparison of Census and AES-FES Estimates for Number of Farms, Area  
and Land Use (in thousands of acres), 1981, Canada<sup>a</sup>

Item	Census Estimate <sup>b, c</sup>	Survey Estimate <sup>c</sup>	Percent Difference <sup>d</sup>
Total number of farms	309,410* <sup>e</sup>	319,476	- 3.2 ± 2.6
Total area of farms	159,866*	175,543	- 8.9 ± 2.4
Improved land	112,390	114,610	- 1.9 ± 2.3
Cropland	75,532*	78,211	- 3.4 ± 2.2
Improved pasture	10,523*	9,460	11.2 ± 7.3
Summerfallow	23,827*	24,939	- 4.5 ± 3.7
Other improved land	2,509*	1,999	25.5 ± 7.7
Unimproved land	47,477*	60,933	- 22.1 ± 4.3
Woodland	8,211*	17,751	- 53.7 ± 3.9
Other unimproved land	39,265*	43,182	- 9.1 ± 6.5

<sup>a</sup> Excluding Newfoundland, Yukon and Northwest Territories.

<sup>b</sup> Excluding specified marginal areas and farms not belonging to the survey universe.

<sup>c</sup> Census and survey totals may not equal the sum of the components due to rounding. Survey estimates for Canada are based on a sample of 18,327 farms.

<sup>d</sup> Percent Difference =  $\frac{(\text{Census Estimate} - \text{Survey Estimate})}{\text{Survey Estimate}} \times 100$ ; the percent difference may not be consistent with the totals represented due to rounding. The indicated confidence interval, resulting from the sampling error in the survey, is equal to  $\pm 2 \times (\text{survey coefficient of variation}) \times \frac{\text{census estimate}}{\text{survey estimate}}$ .

<sup>e</sup> An asterisk, identifying a significant difference between estimates, is indicated when the census estimate lies outside the survey 95 percent confidence interval.

major discrepancies in land area, however, were concentrated in the categories of unimproved land, particularly woodland. Further analysis into the reporting of woodland, which was prompted by these results, is discussed in section 5.5.

Macro-level comparisons also included the study of estimated frequency distributions prepared from the census and survey files. Distributions of the estimated number of farms over variables such as type of organization, land area, area of cropland, and sales were compared. Differences in the distributions identified possible over or undercounting of farms with particular characteristics.

Table 2 presents the census and survey frequency distributions by type of organization for the estimated number of farms in Canada. No significant differences were observed between the estimates for individual or family farms or corporations. However further study was initiated into the coverage of partnerships on the basis of the discrepancies noted for this category.

The limitation of the macro-level comparisons for evaluation of coverage was the inability to separate the effects of response errors from the effects of coverage errors. For example, the differences between census and survey estimates for improved land categories, shown in Table 1, seemed to exhibit too much variation in direction and magnitude to be the result of coverage errors alone. The discrepancies for woodland might also have been caused by factors other than coverage. Perhaps differences in field procedures or questionnaire format had resulted in inconsistencies between the census and surveys in the inclusion or exclusion of land of questionable agricultural value, or the classification of certain categories of land use. The micro-level match provided the needed mechanism for investigating these types of issues.

Table 2  
Comparison of Census and AES-FES Estimates for Number of Farms by  
Type of Organization, 1981, Canada<sup>a</sup>

Type of Organization	Census Estimate <sup>b</sup>	Survey Estimate <sup>c</sup>	Percent Difference <sup>d</sup>
Total number of farms	309,410* <sup>c</sup>	319,476	- 3.2 ± 2.6
Individual or family farm	268,199	267,396	0.3 ± 3.0
Partnership			
- with a written agreement	11,160*	15,908	- 29.8 ± 16.7
- with no written agreement	17,646*	22,855	- 22.8 ± 10.8
Corporation	11,744	12,160	- 3.4 ± 10.4
Other type of organization	661*	1,142	- 42.1 ± 13.0

For footnotes, see Table 1.

5.2 The Micro-level Match

Past experience with other agricultural censuses and surveys has indicated that even the most careful attention to quality cannot entirely prevent response errors. Despite all attempts to provide clear, unambiguous questions, problems such as differing interpretations of certain agricultural terms across regions of Canada, or a lack of consensus on the appropriate classification for certain types of land use, influence the data collected. Misinterpretation is particularly common for items which are of marginal economic or agricultural value, or which do not apply to most respondents. The micro or record level match with the AES-FES files provided the means to evaluate the impact of response errors on the 1981 Census of Agriculture.

The match between the Census of Agriculture and the AES-FES was based on the operator name, address, telephone number, and postal code for each holding. The link was performed in thirteen stages, each requiring a match on a different combination of the identifiers or their components. At each stage of the procedure survey records which has not yet been matched were identified, and the census file was searched by computer to locate the corresponding records. For each survey holding, the specified matching variables or keys were compared character by character with those of the census records which had not yet been linked. A match was identified if all characters of the matching variables were equal. At the Canada level, a computer match rate of 75.7 percent was achieved for the 18,327 survey records.

It was inevitable that, for a certain number of survey records, no census farm would be identified by the computer. Discrepancies in spelling of names and addresses, which had arisen during collection or capture of the census or survey data, prevented links in many cases. For example, J. Smith might have been reported on the census as opposed to J. Smyth on the survey, James Smith as opposed to Jim Smith, or St. Catherines rather than St. Catharines. Partnerships or corporations for which one operator had responded on the census but a different partner or manager had been interviewed by the survey could not be matched by a computer link to operators. Similarly, records for holdings which had changed operators between the census and survey collection dates could not be linked by computer.

In order to improve the match rate, and eliminate the possible biases in the matched file which might have resulted from those operations which could not be linked by computer, a manual resolution process was initiated. Using additional data from the questionnaires, such as corporate or farm name, addresses and names of partners, land description of the holding,

and comments, clerical staff attempted to identify the corresponding census farm for each unmatched survey record. Of the 4,459 unmatched survey farms which remained following the computer link, 3,228 were matched during the manual resolution process. At its completion, 93.3 percent of the total 18,327 AES and FES records for Canada had been linked to census operations.

With further input of time and resources, it may have been possible to link some of the remaining 6.7 percent of the survey records to the census data base. However, in many cases the needed identifiers has not been collected on either the census or survey, and would have required investigation of administrative records or contact with the operators themselves. It was felt that the possible benefits were not sufficient to warrant the expenditures required, and no further manual resolution was attempted.

The studies which were facilitated by the record linkage can be grouped into two main types, those based on the unmatched survey records and those using the matched census-survey pair.

### 5.3 Studies of the Unmatched Records

In order to study the characteristics of census undercoverage, the unmatched records were assumed to be representative of the farms which should have been enumerated by the census but were missed. It was known that the unmatched records overestimated the number of missed farms, due to certain conditions of the data sources and the matching algorithm. For example, it was probable that some records on the survey file had been covered in the census, but could not be matched by either computer or manual means due to missing or invalid name or address data.

Because of the resulting potential for overestimation of the land and commodities missed by the census, one had to proceed with caution in using the estimates produced from the unmatched records. Nonetheless the Canada level estimates were most valuable as initial indicators of the characteristics of the farms which were underenumerated by the census.

In the first stage of the study, sample expansion factors were applied to the unmatched records to produce commodity estimates for the "missed farms". These were then compared to the commodity estimates for the entire survey universe, and the fraction of the total estimate which was accounted for by the "missed farms" was calculated. This fraction was then compared with the fraction which the missed farms comprised of the total estimated number of farms. For example, Table 3 shows that the unmatched file contained only 4.4 percent of the estimated total farm area and 3.9 percent of the cropland, whereas it was responsible for almost 9.7 percent of the total estimate of farms. These results provided an initial indication that the "missed farms" were not representative of the complete universe, but were smaller than average in terms of land area and other characteristics. This implied that the extent of undercoverage could not be measured by the number of farms missed alone. Instead, the characteristics of the missed farms over particular commodities had to be considered.

To provide further insight into the characteristics of undercoverage, frequency distributions of the estimated number of missed farms were prepared over classes of land area, sales, livestock, and other commodities. Comparison with similar frequency distributions for the entire survey universe indicated that the missed farms had a higher proportion of holdings with small acreages and low sales. It can be seen from Table 4, for example, that 42.3 percent of the estimated missed farms reported less than 70 acres of total farm area, as compared with 15.8 percent of the complete survey population. Sales of less than \$1,200 were reported by an estimated 27.7 percent of the missed farms, but only 7.3 percent of the survey universe.

The frequency distributions were also compared by considering the ratio of the unmatched estimate to the estimate for the entire survey universe, that is, the fraction of the total survey estimate accounted for by the unmatched farms. As shown in Table 4, the unmatched file contained 36.5 percent of the estimated farms with less than 10 acres of land, the smallest size range presented as compared with only 4.3 percent of those in the largest range of 760 acres or more. Similarly 36.8 percent of the estimated farms with sales less than \$1,200 were obtained

Table 3

Comparison of AES-FES Estimates for Total Farms and Unmatched Farms,  
Area and Land Use (in thousands of acres), 1981, Canada<sup>a</sup>

Item	Estimate from Total AES-FES File <sup>b</sup>	Estimate from Unmatched AES-FES Records <sup>c</sup>	Percent of the Total Estimate Accounted for by the Un- matched Farms
Number of Farms	319,476	30,975	9.7
Total Farm Area	175,543	7,768	4.4
Total Improved Land	114,610	4,502	3.9
Cropland	78,211	2,792	3.6
Improved Pasture	9,460	603	6.4
Summerfallow	24,939	1,004	4.0
Other Improved Land	1,999	104	5.2
Total Unimproved Land	60,933	3,266	5.4
Woodland	17,751	1,325	7.5
Other Unimproved Land	43,182	1,941	4.5

<sup>a</sup> Excluding Newfoundland, Yukon and Northwest Territories.

<sup>b</sup> Survey estimates are based on a sample of 18,327 farms.

<sup>c</sup> The unmatched file contained 1,231 farms.

Table 4

Percentage Distribution of AES-FES Estimates for Total Farms  
and Unmatched Farms by Total Farm Area and Total Value of Agricultural  
Products Sold During 1980, 1981, Canada<sup>a</sup>

Item	AES-FES Estimate of Total Number of Farms <sup>b</sup>	AES-FES Estimate of Number of Unmatched Farms <sup>c</sup>	Percent of the Total Estimate Accounted for by the Unmatched Farms
	Cumulative Percent	Cumulative Percent	Percent
Total Farm Area			
Under 10 acres	3.5	13.2	36.5
10 - 69 acres	15.8	42.3	22.9
70 - 399 acres	62.8	85.0	8.8
400 - 759 acres	78.4	90.5	3.4
760 acres and over	100.0	100.0	4.3
Total Value of Agricultural Products Sold			
Under \$1,199	7.3	27.7	36.8
\$ 1,200 - \$ 2,499	12.3	41.2	26.5
2,500 - 9,999	29.6	65.4	13.5
10,000 - 49,999	67.8	88.3	5.8
50,000 and over	100.0	100.0	3.5

<sup>a</sup> Excluding Newfoundland, Yukon and Northwest Territories.

<sup>b</sup> Survey estimates are based on a sample of 18,327 farms.

<sup>c</sup> The unmatched file contained 1,231 farms.



from the unmatched file, compared to 3.5 percent of those with sales of \$50,000 or more.

In summary, the results of the study of unmatched records were able to provide concrete evidence that the holdings missed by the census tended to be smaller than average in terms of agricultural production and value, a theory which had been widely held but not proven.

#### 5.4 Studies Using the Matched File

The matched file was composed of agricultural holdings for which census and survey records could be linked by the computer and manual processes described. Since these census and survey values were assumed to have been collected from the same set of holdings, the effects of coverage differences were removed. In addition, the influence of imputation was lessened by excluding records for which census or survey data had been entirely imputed due to non-response. Hence the nature and extent of potential response differences between the two data collection vehicles could be studied. Prior to discussing the results of the matched record studies, however, a number of limitations which existed within the matched file, and which influenced the evaluation process, should be described.

Although every effort was taken to lessen the chance of spurious matches, it is possible that a small number of survey farms may have been linked to the wrong census holding due to similarities in name or address. In the case of extremely large agricultural operations which made a significant contribution to provincial commodity totals or land areas, linkage to the wrong census operation could noticeably skew the results. A detailed study of a sample of matched records, undertaken to determine the quality of the computer link, had not been completed at the time of the evaluation. As a result, the potential influence of spurious matches had to be considered when studying results from the matched file.

The second limitation on the matched file analysis affected the comparison of total counts from the census and survey data. The matched records consisted of a subset of the non-self-weighting sample of the AES and FES, since they contained only the survey farms which could be linked to census records. It would have been preferable to apply sample expansion factors to produce weighted estimates for the matched file. However, the expansion factors for the area frame were calculated using reported land use values from the survey, and it was as yet undetermined whether the factors were valid when applied to census data from the matched file. Census-related expansion factors could not be calculated using the census data, since one component of the factor, the farm area inside the selected segment of land, was collected only by the survey. As a result of this uncertainty regarding the application of survey expansion factors to census matched data, it was decided to restrict the analysis to unweighted census and survey totals from the matched file. (Study of the use of weighted estimates from the matched file was underway at the time of writing, but no conclusions had yet been reached.)

Despite the limitations of a non-self-weighting sample and the possible existence of spurious matches, the matched file proved to be a valuable evaluation tool. When matched totals identified discrepancies between census and survey values, further detailed investigations were undertaken into the possible causes of the observed differences.

As an example of the use of the unweighted matched totals, Table 5 presents counts for total farm area and categories of land use at the Canada level. The results indicate that less land was reported on the census than the survey for all land use items except improved pasture and other improved land. Relative differences between census and survey totals were smallest for items such as cropland, which are of major economic value and hence are clearly defined and seldom misunderstood by farm operators. Items of more marginal agricultural and economic value, however, tended to display greater discrepancies. The largest relative differences were observed for the category of woodland. In order to demonstrate some of the detailed evaluation techniques made possible by the matched file, further results of the study of data on woodland will be discussed.

**Table 5**  
Comparison of Census and AES-FES Totals for Matched Farms,  
Land Use (thousand of acres), 1981, Canada<sup>a</sup>

Item	Census Total <sup>b</sup>	Survey Total <sup>b</sup>	Percent Difference <sup>c</sup>
Total area of farms	13,059	14,091	- 7.3
Improved land	8,798	8,801	- -
Cropland	6,046	6,167	- 1.9
Improved pasture	804	682	18.0
Summerfallow	1,777	1,816	- 2.1
Other improved land	170	137	24.3
Unimproved land	4,261	5,291	- 19.5
Woodland	523	1,102	- 52.5
Other unimproved land	3,737	4,189	- 10.8

<sup>a</sup> Excluding Newfoundland, Yukon and Northwest Territories.

<sup>b</sup> Records for which census or survey data were entirely imputed have been excluded leaving 16,388 matched farms. Census and survey totals may not equal the sum of the components due to rounding.

<sup>c</sup> Percent Difference =  $\frac{\text{Census total} - \text{Survey total}}{\text{Survey Total}} \times 100$

### 5.5 Detailed Comparisons of the Matched Records

While the matched totals provided a measure of the overall biases in reporting, they masked detailed information on where and why the differences occurred. For instance, was the difference in woodland caused by large discrepancies in only a handful of holdings, or was it consistent across all records? Did the response differences vary in magnitude or direction across types of operations or regions of the country? By comparing census and survey responses at the individual record level, the characteristics of the reporting differences were studied in detail.

Table 6 presents an example of the type of investigation facilitated by the matched evaluation file. Only those holdings for which a non-zero value of woodland was reported on either the census or survey are included. The operations are classified by the size and direction of the difference between the census and survey values for the item, and cross-classified by the amount of woodland reported on the census.

The table indicates that less woodland was reported on the census than on the survey for the majority of holdings. Differences tended to be small, clustered in the 1 to 50 acres and 51 to 150 acres ranges even for operations with large amounts of woodland on the census. Of note also were the 3,177 holdings, or 34.3 percent of the universe of interest, for which the census value of woodland was zero but the survey value was greater than zero. This is in contrast to the 807 holdings, or 8.7 percent of the universe, in which the opposite case of zero acres of woodland on the survey but greater than zero acres on the census was observed. Examination of these results suggested that the census and surveys may not have been obtaining measures of the same quantity, and prompted further study into their collection methodologies and questionnaire formats.

On the census questionnaire the respondent was asked to report the area of woodland, with further instructions in the census representative's manual indicating that only land "with seedlings or trees which had or would have value as timber, fuelwood, or Christmas trees" be included. In contrast the AES and FES interviewers instructed respondents to "include woodlots, cut-over land, etc." with no additional instructions that the land be of present or future commercial value. As well, woodland was requested twice on the surveys, once immediately after the reporting of total area, and later as a component of land use, and thus received greater emphasis.



**Table 6**  
Comparison of Census and AES-FES Responses for Matched Records,  
Difference in Woodland by Census Value of Woodland, Canada<sup>a</sup>, 1981

Difference:	Census Value of Woodland (acres)							Total	Percent of Re- porting Farms
Census Woodland - Survey Woodland (acres)	0	1 to 2	3 to 9	10 to 69	70 to 239	240 to 399	400 or more		
Number of Reporting Farms <sup>b</sup>									
less than - 500	241	0	3	16	16	6	11	293	3.2
- 500 to - 251	248	0	4	24	21	12	7	316	3.4
- 250 to - 151	268	4	1	27	44	9	4	357	3.9
- 150 to - 51	715	5	30	167	145	24	19	1,105	11.9
- 50 to - 1	1,705	59	277	1,053	370	57	28	3,549	38.3
0	-	26	165	481	151	21	21	865	9.3
1 to 50	-	55	229	1,288	492	53	30	2,147	23.2
51 to 150	-	-	-	50	294	45	32	421	4.5
151 to 250	-	-	-	-	65	25	17	107	1.2
251 to 500	-	-	-	-	-	33	42	75	0.8
greater than 500	-	-	-	-	-	-	36	36	0.4
Total	3,177	149	709	3,106	1,598	285	247	9,271	100.0

<sup>a</sup> Excluding Newfoundland, Yukon, and Northwest Territories.

<sup>b</sup> Including all operations which reported woodland on either the census or survey. Excluding records for which either the census or survey data were totally imputed due to non-response.<sup>c</sup>

As a result of these differences, it was believed that certain areas of woodland of questionable commercial value, which were reported on the survey, may have been excluded from the census. As well, some areas may have been reported on the census under different categories of land use, such as other unimproved land. Study continues into these and other hypotheses, using the matched file to investigate possible causes of the observed response differences.

When summary tabulations from the matched file failed to suggest causes for observed biases, the study of individual records which displayed large discrepancies between census and survey values for the items of interest was often informative. By comparing census and survey responses for other related items, it was sometimes possible to identify misclassification between categories or other causes of reporting differences.

Table 7 shows a number of variables for a record with one of the largest differences between census and survey values for total farm area. In this case, the discrepancy in total area was

**Table 7**  
Census and Survey Recorded Values for a Particular Matched Record

	Total Farm Area (acres)	Land Use (acres)						Total Cattle and Calves
		Crop- land	Improved Pasture	Summer- fallow	Other Improved Land	Wood- land	Other Unim- proved Land	
Census Values	2,640	1,035	0	1,240	15	0	350	920
Survey Values	17,000	1,010	0	970	20	0	15,000	815

concentrated in the category of other unimproved land; only minor differences existed between responses for cropland, summerfallow, and other improved land. It appears that most of the land which was reported as other unimproved on the survey was excluded from the census response. Similar studies of other individual records identified other potential discrepancies in the reporting of unimproved land. Theories developed from these studies of individual cases were then tested on the entire file to determine if they might apply in general.

## 6. CONCLUSION

The link between the 1981 census and survey files was a powerful tool for the evaluation of census coverage and response errors. Errors were known to exist in the survey data which were being used as approximations to the true values, and limitations of the linkage operation were known to have caused spurious matches and mismatches. However, the matched file was still a valuable source of data for investigation of quality concerns. Studies based on the record level matches broadened the initial results obtained from the macro-comparison of census and survey estimates. In addition they brought individual problems into focus by allowing detailed investigation of particular aspects.

The evaluation produced valuable results on census undercoverage. Studies based on the unmatched survey records showed that the holdings missed by the census were, in general, smaller than average in terms of total land area, livestock, and value of agricultural products sold. Thus concrete evidence was provided to support the widely-held theory that the census tended to miss holdings of marginal economic and agricultural value.

Studies of response differences for land use identified categories in which discrepancies were concentrated, tendencies for confusion between certain classes, and variations in differences among regions of the country. Possible revisions to questionnaire format and wording, or collection methods, have been considered as a result of the study. Some of the variations are known to have been caused by difficulties in defining certain land use categories, due to the lack of clarity in the concepts themselves. Problems such as these, that result from confusion in the minds of the respondents as to which land should be reported under which category of usage, may never be completely solved. However, the recognition of the existence of a problem, and the study of the characteristics of its occurrence and its effect on the data, are very valuable contributions to future planning.

The 1981 data quality evaluation had far-reaching effects for the census. In response to its main goal, the study identified quality concerns in the 1981 census data. A publication (Statistics Canada 1984) was prepared to provide users with an indication of data quality, and to advise them with respect to particular problems which had a noticeable impact on the data. Looking further ahead, the evaluation has served as input to the planning of 1986 census procedures. By identifying items for which coverage or response errors occurred in 1981, the study has provided a list of areas requiring further consideration of collection and processing methods.

The impact of the data quality evaluation was not restricted to the Census of Agriculture alone. The comparison of census and survey responses also identified problem areas in the other data collection vehicles. Improvements to the National Farm Survey, the annual probability survey which has replaced the AES and FES, may result from the census study.

A further benefit of the study is the knowledge gained on the use of record linkage for evaluation purposes. The experience in matching data at the individual record level, using both computer and manual means, could provide valuable input to other linkage projects. In particular, knowledge of the problems encountered, their causes, characteristics, and possible solutions, could result in improved procedures for other studies.

In summary, the 1981 Census of Agriculture Data Quality Evaluation project has provided further evidence of the power of matching in the study of non-sampling error. The investigations using matched and unmatched records, and macro and micro-level comparisons, have

produced measures of quality of the 1981 census data, and identified items for which error have impacted significantly upon the results. As an extension of the original project mandate input was provided to the planning of the 1986 and subsequent censuses, by indicating areas in which further research into possible changes in methodology was required. The evaluation also identified potential problems in the surveys used for comparison, thereby contributing to the planning of future vehicles for the collection of agricultural data. Finally, the study provided valuable experience and insight into the application of record linkage techniques for data quality evaluation.

### ACKNOWLEDGEMENTS

The author would like to thank D. Royce, M.R. Dibbs, B.N. Chinnappa, K. Thatcher, the reviewer, and the other members of Statistics Canada whose support and valuable comments were most appreciated during the writing of this paper.

### REFERENCES

- ANDERSEN, R., KASPER, J., FRANKEL, M.R., and associates (1979). *Total Survey Error*. San Francisco: Jossey-Bass Publishers.
- COCHRAN, W.G. (1977). *Sampling Techniques*, third edition. New York: John Wiley & Sons.
- DEMING, W.E. (1944). On Errors in Surveys. *American Sociological Review*, 9, 359-369.
- FAULKENBERRY, D., and TORTORA, R.D. (1981). Non-sampling Errors in an Agriculture Survey. *1981 Proceedings of the Section on Survey Research Methods, of the American Statistical Association*, 493-495.
- FELLEGI, I.P. (1964). Response Variance and its Estimation. *Journal of the American Statistical Association*, 59, 1016-1041.
- FELLEGI, I.P. (1973). The Evaluation of the Accuracy of Survey Results: Some Canadian Experience. *International Statistical Review*, 41, 1-14.
- GOSSELIN, J.-F., CHINNAPPA, B.N., GHANGURDE, P.D., and TOURIGNY, J. (1978). *A Compendium of Methods of Error Evaluation in Censuses and Surveys* (Catalogue 13-564). Ottawa, Canada: Statistics Canada.
- HANSEN, M.H., HURWITZ, W.N., MARKS, E.S., and MAULDIN, W.P. (1951). Response Error in Surveys. *Journal of the American Statistical Association*, 46, 147-190.
- HANSEN, M.H., HURWITZ, W.N., and BERSHAD, M. (1961). Measurement Errors in Censuses and Surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.
- HARTLEY, H.O. (1981). Estimation and Design for Non-sampling Errors of Surveys. In *Current Topics in Survey Sampling*, ed. D. Krewski, R. Platek, and J.N.K. Rao. New York: Academic Press.
- HORVITZ, D.G. (1981). Response Error Research Issues in Health Surveys. *1981 Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 397-399.
- KIBLER, W.E. (1978). Controlling Non-sampling Errors in Surveys. Summary Report of the 29<sup>th</sup> Federal Provincial Committee on Agricultural Statistics, Statistics Canada.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.
- KISH, L., and LANSING, J.B. (1954). Response Errors in Estimating the Value of Homes. *Journal of the American Statistical Association*, 49, 520-538.
- KROTKI, K. (1980). *Response Error in the 1976 Census of Population and Housing*. Working Paper. Ottawa, Canada: Minister of Supply and Services Canada.
- MARQUIS, K.H., MARQUIS, M.S., and POLICH, J.M. (1981). Survey Responses to Sensitive Topics. *1981 Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 339-341.

- NISSELSON, H., and BAILLAR, B.A. (1976). Measurement, Analysis, and Reporting of Nonsampling Errors in Surveys. *Proceedings of the 9th International Biometric Conference*, 2, 201-322.
- PHILLIPS, J. (1978). 1979 Farm Expenditure Survey Design and Estimation Procedures. Working Paper, Institutional and Agriculture Survey Methods Division, Statistics Canada.
- STATISTICS CANADA (1979). *1976 Census of Canada - Agriculture - Evaluation of Data Quality* (Catalogue 96-872). Ottawa, Canada: Minister of Supply and Services Canada.
- STATISTICS CANADA (1980). *1976 Census of Canada - Quality of Data - Series I: Sources of Error - Coverage* (Catalogue 99-840). Ottawa, Canada: Minister of Supply and Services Canada.
- STATISTICS CANADA (1982). *1981 Census of Canada - Agriculture* (Catalogue 96-901). Ottawa, Canada: Minister of Supply and Services Canada.
- STATISTICS CANADA (1984). *1981 Census of Canada - Agriculture - Evaluation of Data Quality* (Catalogue 96-918). Ottawa, Canada: Minister of Supply and Services Canada.
- SUKHATME, P.V., and SETH, G.R. (1952). Non-sampling Errors in Surveys. *Journal of the Indian Society of Agriculture Statistics*, 4, 5-41.
- TREMBLAY, V., SINGH, M.P., and CLAVEL, L. (1976). Methodology of the Labour Force Survey Re-interview Program. *Survey Methodology Journal*, 2, 43-62.
- U.S. BUREAU OF THE CENSUS (1964). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Accuracy of Data on Population Characteristics as Measured by the CPS - Census Match*. Series ER60, No. 5, Washington, D.C.: U.S. Government Printing Office.
- U.S. BUREAU OF THE CENSUS (1970). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Record Check Study of the Accuracy of Income Reporting*. Series EDR60, No. 8, Washington, D.C.: U.S. Government Printing Office.
- U.S. BUREAU OF THE CENSUS (1982). *1978 Census of Agriculture Volume 5 Special Reports - Part 3 Coverage Evaluation*, (AC78-SR-3). Washington, D.C.: U.S. Government Printing Office.





## Unbiased Estimation of Domain Parameters in Sampling without Replacement

ARIJIT CHAUDHURI AND RAHUL MUKERJEE<sup>1</sup>

### ABSTRACT

A finite population of size  $N$  is supposed to contain  $M$  (unknown) units of a specified category  $A$  (say) constituting a domain with mean  $\mu$ . A procedure which involves drawing units using simple random sampling without replacement till a preassigned number of members of the domain is reached is proposed. An unbiased estimator of  $\mu$  is also derived. This is seen to be superior to the corresponding possibly biased estimator based on a comparable SRSWOR scheme with a fixed number of draws. The proposed scheme is also shown to admit unbiased estimators of  $M$  and the domain total  $T$ .

KEY WORDS: Domain estimation; Simple random sampling without replacement.

### 1. INTRODUCTION

In large scale sample surveys, utilization of available resources and consideration for efficiency often demand realization in a sample of adequate representation from a specified category ( $A$ , say) of members with required characteristics. For example, clients and users of survey data may insist on estimates from a sample with a specified ( $m$ , say) number: (1) of farmers (i) using a particular fertilizer, (ii) employing a particular irrigation and cultivation technique and (iii) ready to respond truthfully to queries made; (2) of manufacturers using iron and steel with a specific purpose; (3) of household members with a requisite academic qualification, etc. While designing a sampling plan for the purpose, in spite of careful efforts, it is often possible that 'frames' may not be accurately constructed. The faulty list may be supposed to include  $N$  units which are well in excess over the  $M$  genuine units of the required  $A$ -category. Hence arises a problem of sampling to yield estimators for the mean, (and also total and size), of the domain of  $A$ -members. A solution to this problem is attempted below using 'inverse' SRSWOR scheme. Inverse sampling plans with replacement are, however, available in the literature (vide Haldane 1945, Sampford 1962 among others) for estimating the proportion  $f = M/N$  of domain elements. Domain estimators for  $\mu$  are also given by Rao (1975) but they are ratio estimators and are not unbiased. The proposed inverse SRSWOR scheme is seen to admit an unbiased estimator of  $\mu$  which is more efficient than the corresponding possibly biased estimator based on a comparable SRSWOR scheme with the fixed number of draws.

### 2. A METHOD OF SAMPLING AND ESTIMATION

The population  $I_N = (1, \dots, j, \dots, N)$  is supposed to consist of  $N$  units labelled  $1, \dots, j, \dots, N$  and valued  $y_1, \dots, y_j, \dots, y_N$ . Of them some  $M$  (unknown) units possess certain exclusive features or constitute a class or domain, say  $A$ . In practice, some idea about  $M$  is usually available and let the parameter space for  $M$  be  $\mathcal{M} = \{r, r+1, \dots, R\}$ , where  $r(\geq 1)$  and  $R(\leq N)$  are known. In almost all real life situations  $r$  will be much greater than 1 and  $R$  much less than  $N$ .



Writing  $X_1, \dots, X_i, \dots, X_M$  as the  $y_j$  values for the  $M$  units of  $A$ , estimators are required for  $\mu = (\sum_1^M X_i)/M$ , and perhaps also for  $M$  and  $T = \sum_1^M X_i$ , from a sample containing a preassigned number, say  $m$  ( $\leq r$ ), of units of  $A$ . The expressions for the variances of these estimators, presented later as functions of  $m$ , may be employed for an appropriate choice of  $m$ . For convenience, we shall write  $X_{M+1} = \dots = X_N = 0$  for the 'non- $A$ ' units of  $I_N$ .

Let units be chosen in successive draws by SRSWOR till exactly  $m$  units of  $A$  are realized. The number of draws,  $u$ , is then a random variable with a probability distribution  $P_M(\cdot)$  (say, depending on the unknown parameter  $M$ ) which is given by

$$P_M(u=n) = \frac{\binom{M}{m-1} \binom{D}{n-m}}{\binom{N}{n-1}} \cdot \frac{M-m+1}{N-n+1} = g_{Mn} \text{ (say) } (m \leq n \leq D+m), \quad (2.1)$$

where  $D = N - M$ . To avoid trivialities, hereafter, we shall make the reasonable assumption that  $m \geq 2$ . Then the following results hold for the above inverse sampling scheme.

**Lemma 2.1.** Every parametric function  $f(M)$  is unbiasedly estimable.

**Proof.** Let  $h(u)$ , if available be an unbiased estimator (UE) of  $f(M)$ . Then

$$f(M) = \sum_{n=m}^{D+m} h(n) g_{Mn}, \quad r \leq M \leq R. \quad (2.2)$$

If the above system of  $R - r + 1$  equations in  $N - r + 1$  unknowns  $h(m), \dots, h(N - r + m)$  be written in matrix notation, then the fact that  $g_{Mn} > 0$  ( $m \leq n \leq D + m, r \leq M \leq R$ ) implies that the resulting coefficient matrix is of full row rank. This guarantees the existence of a solution and completes the proof.

**Remark.** In particular, if  $R = N$  then the number of equations in (2.1) equals the number of unknowns. As such the coefficient matrix becomes nonsingular and every parametric function  $f(M)$  becomes uniquely unbiasedly estimable.

**Corollary 2.1.** A UE of  $M$  based on  $u$  is  $\hat{M}(u) = N(m-1)/(u-1) = \hat{M}$  (say).

**Proof.** First observe that the assumption  $m \geq 2$  ensures that  $u > 1$  with probability 1 (whatever be  $M$ ) so that  $\hat{M}(u)$  is well defined. Now

$$\begin{aligned} E\left(\frac{1}{u-1}\right) &= \sum_{n=m}^{D+m} \frac{1}{n-1} g_{Mn} \\ &= \frac{M! D!}{(M-m)! (m-1)! N!} \sum_{n=m}^{D+m} \frac{(n-2)! (N-n)!}{(n-m)! (D-n+m)!} \\ &= \frac{M! D!}{(M-m)! (m-1)! N!} \cdot \frac{(M-m)! (m-2)! (N-1)!}{(M-1)! D!} = \frac{M}{N(m-1)}, \quad \forall M \in \mathcal{M}. \end{aligned}$$

Hence the result.

**Remark.** The relation (2.1) and Lemma 2.1 may be employed to find  $V_M(\hat{M})$  and a UE of this variance. The resulting algebraic expression, although straightforward to evaluate numerically in any practical situation, are somewhat involved and will not be presented here.

In the following,  $S^2 = (M-1)^{-1} \sum_1^M (x_i - \mu)^2$ ,  $q(u)$  and  $\ell(u)$  are any UE's for  $M^{-1}$  and  $M^2$  respectively (available by (2.2) above)  $\Sigma'$  denotes summation over the  $A$ -units included in the sample,  $\bar{x} = m^{-1} \Sigma' X_i$ ,  $Z = m^{-1} \Sigma' X_i^2$  and  $s^2 = (m-1)^{-1} \Sigma' (X_i - \bar{x})^2$ .

**Theorem 2.1.** A UE of  $\mu$  is  $\bar{x}$  with  $V_M(\bar{x}) = S^2(1/m - 1/M)$ . A UE of  $V_M(\bar{x})$  is given by  $v(\bar{x}) = s^2(m^{-1} - q(u))$ .

**Proof.** Easy and hence omitted.

**Theorem 2.2.** (i) A UE of  $T$  is  $\hat{T} = \hat{M}\bar{x}$  with

$$V_M(\hat{T}) = S^2(1/m - 1/M) E_M(\hat{M}^2) + \mu^2 V_M(\hat{M}).$$

(ii)  $\nu(\hat{T}) = \hat{T}^2 - [\ell(u)(Z - s^2) + \hat{M}s^2]$  is a UE of  $V_M(\hat{T})$ .

**Proof.** The proof of (i) is easy and hence omitted. To prove (ii) note that

$$\begin{aligned} &E [ \{ \ell(u)(Z - s^2) + \hat{M}s^2 \} \mid u ] \\ &= \ell(u)(M^{-1} \sum_1^M X_i^2 - S^2) + \hat{M}s^2 = \ell(u)(\mu^2 - M^{-1}S^2) + \hat{M}s^2. \end{aligned}$$

Hence

$$E_M \nu(\hat{T}) = E_M(\hat{T}^2) - [M^2(\mu^2 - M^{-1}S^2) + MS^2] = E_M(\hat{T}^2) - T^2 = V_M(\hat{T}).$$

3. COMPARISON WITH SRSWOR WITH A FIXED NUMBER OF DRAWS

In this section, first it will be shown that if one insists on unbiased estimation of  $\mu$  then our strategy will be superior to the one based on SRSWOR with a fixed number of draws. Secondly, this superiority will be demonstrated even when biased estimators are allowed.

Let  $d$  be a fixed (somehow) number of draws in SRSWOR sampling,  $\hat{s}$  a sample so drawn,  $\cap A$  the set of  $A$ -units in  $\hat{s}$  and  $C$  the cardinality of  $\hat{s} \cap A$ . We will use, for this scheme also previous notations  $P_M, E_M, V_M$  to imply phenomena relevant here. Then for such a sampling we have:

**theorem 3.1.**  $\mu$  admits a UE if and only if  $d \geq N - r + 1$ .

**roof.** Let  $d \geq N - r + 1$ . Then  $P_M [c = 0] = 0, \forall M \in \mathcal{M}$  and  $\hat{\mu} = c^{-1} \Sigma' X_i$  is a UE of  $\mu$ . To prove the necessity it will be enough to show that if  $d = N - r$ , then  $\mu$  does not admit a UE. For this the following notations will be used. let  $j_1, \dots, j_d$  be  $d$  distinct increasingly ordered units out of  $1, \dots, N$ , constituting the elements of  $\hat{s}$  and such that some  $k$  of them ( $0 \leq k \leq d$ ),  $y_{i_1}, \dots, i_k$  (increasingly ordered) belong to  $A$ . Then we write  $\hat{s} = (j_1, \dots, j_d), \hat{s}' = (i_1, \dots, i_k) \cap A$  (so that  $k = 0 \Rightarrow \hat{s}' = \Phi$  and  $k = d \Rightarrow \hat{s}' = \hat{s}$ ) and  $X(\hat{s}') = (X_{i_1}, \dots, X_{i_k})$ , a quence of  $X_i$  values for the units in  $\hat{s}'$ . Then if there exists a UE for  $\mu$ , say  $t$ , we may write  $= t(X(\hat{s}')|\hat{s})$  such that

$$E_M(t) = \mu, \forall X_1, \dots, X_M \forall M \in \mathcal{M}. \tag{3.1}$$

For  $0 \leq k \leq d$ , let  $t_k = \Sigma_k t(X(\hat{s}')|\hat{s})$ ,  $\Sigma_k$  being sum over all samples with exactly  $k$   $A$ -units. Clearly  $t_0$  is free from  $X_i$ 's.

If  $d = N - r$ , then  $\mathcal{M} = \{N - d, N - d + 1, \dots, N\}$ . Suppose  $M = N - d + j$  ( $1 \leq j \leq d$ ). Then the  $A$ -units may be chosen in  $\binom{N}{j} = \binom{N-d}{d-j}$  ways. Accordingly  $\binom{N-d}{d-j}$  equations are involved in (3.1). Summing over the number of ways of choosing the  $A$ -units, (3.1) yields

$$\sum_{w=j}^d a_{jw} t_w = \frac{\binom{N}{d} \binom{N-1}{d-j} T}{N - d + j}, \tag{3.2}$$

where, for  $0 \leq j \leq w \leq d$ ,  $a_{jw} = \binom{N-d}{w-j}$  if  $N-d \geq w-j$ ; 0, otherwise. From (3.2) the solutions for the  $t_w$ 's may be obtained as

$$t_w = \binom{N}{d} \binom{d}{w} \frac{T}{N}, \quad 0 \leq w \leq d, \quad (3.3)$$

and the validity of (3.3) follows from the fact that

$$\sum_{w=j}^d \binom{N-d}{w-j} \binom{d}{w} = \binom{N}{d-j}.$$

In particular, (3.3) yields  $t_0 = N^{-1} \binom{N}{d} T$ . But then  $t_0$  is not free from the  $X_i$ 's implying a contradiction, proving the necessity and completing the proof.

Thus with a fixed size ( $d$ ) SRSWOR scheme, for unbiased estimation of  $\mu$  we need  $d \geq N - r + 1$  which may become too large (especially if  $r$  is small) making the scheme operationally inconvenient. Even if  $d \geq N - r + 1$ , the fixed size SRSWOR scheme together with the UE  $\hat{\mu} = c^{-1} \Sigma' X_i$  can be seen to be less efficient than the strategy described in the preceding section when compared at equal level of cost of inspection.

To elaborate, suppose  $d \geq N - r + 1$  and note that

$$V_M(\hat{\mu}) = S^2 \left[ E_M(1/c) - 1/M \right]. \quad (3.4)$$

For our inverse sampling scheme, by (2.1) the expected number of draws is given by  $m(N+1)/(M+1)$  and, to make our scheme comparable to a fixed size ( $d$ ) scheme, this should equal  $d$  i.e. one should have  $m = d(M+1)/(N+1)$ , in which case Theorem 2.1 yields

$$V_M(\bar{x}) = S^2 \left[ \frac{N+1}{d(M+1)} - \frac{1}{M} \right]. \quad (3.5)$$

Since

$$E_M(c^{-1}) > [E_M(c)]^{-1} = \frac{N}{dM} > \frac{N+1}{d(M+1)},$$

it follows that (3.4) is greater than (3.5), proving our assertion.

It is also interesting to compare our strategy with the fixed size scheme when a possibly biased estimator of  $\mu$  is allowed in the latter. In fixed size ( $d$ ) SRSWOR scheme, consider the usual (ratio) estimator of  $\mu$  given by [vide e.g. Rao (1975)]

$$\begin{aligned} \mu^* &= c^{-1} \Sigma' X_i & \text{if } c > 0 \\ &= 0 & \text{if } c = 0 \end{aligned}$$

The bias in  $\mu^*$  equals  $-\mu P_M(c=0)$  (observe that if  $d \geq N - r + 1$ , then  $P_M(c=0) = 0$ ,  $\forall M \in \mathcal{M}$  and  $\mu^*$  reduces to the UE  $\hat{\mu}$  defined earlier) and it can be shown that

$$MSE_M(\mu^*) = S^2 \sum_{a \geq 1} (1/a - 1/M) P_M(c=a) + \mu^2 P_M(c=0). \quad (3.6)$$

A straightforward analytic comparison between (3.5) and (3.6) is difficult but as numerical examples including the two cited below suggest, in most practical situations (3.5) will be smaller than (3.6), indicating the superiority of our strategy even when a possibly biased estimator is allowed in the fixed size scheme.

**Example 3.1.** The following data relate to the aggregate percentage of marks of all the students who passed the Bachelor of Statistics Examination of the Indian Statistical Institute (ISI) during the last five academic years ended 1984<sup>1</sup>.

<sup>1</sup> The data are obtained from the office of the ISI Dean of Studies to whom the authors are grateful for granting an access to them.

68	80	80	72	87	71	55	75	85	52	82
76	73	54	57	51	56	48	73	54	76	69
87	81	68	74	58	56	71	66	69	81	59
65	83	79	72	50	44	65	61	57	50	73
85	87	64	70	48	58	61	53	56	62	61
74	62	56	62	58	58	66	70	80	74	80

Suppose it is desired to estimate from a sample the mean score of those students who obtained a first class (i.e. sixty percent or above). Then  $N = 66$ ,  $M = 44$ ,  $\mu = 73.1818$ ,  $S^2 = 61.6871$ . For a fixed size SRSWOR scheme with  $d = 10$ , (3.6) equals 9.5967. The comparable  $m$  in our inverse sampling strategy is  $d(M + 1)/(N + 1) = 6.72$  and with this  $d = 6, 7$ , (3.5) equals 8.8792, 7.4105 and the resulting gains in efficiency, compared to the fixed size scheme, are 8.08 and 29.50 percent respectively.

**Example 3.2.** As a somewhat less traditional example, consider the problem of estimating the mean of the prime numbers among the first sixty natural numbers. The  $N = 60$ ,  $M = 18$ ,  $\mu = 24.5$ ,  $S^2 = 350.1471$ . For a fixed size SRSWOR scheme with  $d = 7$ , the value of (3.6) is 205.4654. The comparable inverse sampling strategy requires  $m = d(M + 1)/(N + 1) = 2.18$  and with this  $d = 2$ , (3.5) equals 155.6209, indicating a gain in efficiency by 32.03 percent.

#### 4. CONCLUDING REMARKS

In this paper we have considered the estimation problem for a single domain. In large scale surveys estimators are often required for several domains in which case the present procedure may be modified as follows.

Let there be  $t$  domains, the domain sizes  $M_k$  being unknown, having respective parameter spaces  $\mathcal{M}_k = \{r_k, r_{k+1}, \dots, R_k\}$ , where  $r_k$  and  $R_k$  are known ( $1 \leq k \leq t$ ). Let  $\mu_k$  and  $S_k^2$  denote the population mean and variance of the study variate in the  $k$ th domain. The sampling scheme may be inverse generalized hypergeometric, i.e. inverse SRSWOR may be continued till at least  $m_1, m_2, \dots, m_t$  ( $m_k \leq r_k$  for each  $k$ ) units of the 1st, 2nd, ...,  $t$ th domains are realized. For each  $k$ , clearly the number of units, say  $\xi_k$ , in the sample from the  $k$ th domain is now a random variable, with  $P_{1r}(\xi_k \geq m_k) = 1$  (where  $m = (M_1, \dots, M_t)'$ ,  $P_{1r}$ ,  $E_{1r}$  the corresponding probability and expectation operator), since even when the quota for  $k$ th domain is filled up, sampling may have to be continued to fill up those for the other domains thus possibly including in the sample some additional units from the  $k$ th domain. The mean,  $\bar{x}_k$ , of the units in the sample from the  $k$ th domain is a UE of  $\mu_k$  with a variance  $S^2 [E_{1r}(\frac{1}{\xi_k}) - \frac{1}{M_k}]$ ,  $1 \leq k \leq t$ .

In this set-up also numerical investigations (records omitted since they seem uninteresting in the present context) suggest that the inverse sampling strategy will be more efficient than the one based on fixed size SRSWOR when compared at the same level of cost. For multidomain situations, however, the detailed algebraic expressions become somewhat involved so that an analytic comparison along the line of the preceding section becomes difficult and hence not reported here.

#### ACKNOWLEDGEMENT

The authors are thankful to the referee for his highly constructive suggestions that helped improvement on an earlier draft.

#### REFERENCES

- Haldane, J.B.S. (1945). On a method of estimating frequencies. *Biometrika*, 33, 222-225.
- Rao, J.N.K. (1975). Analytical studies of sample survey data. *Survey Methodology*, 1, 1-76.
- Sampford, M.R. (1962). Methods of cluster sampling with and without replacement for clusters of unequal sizes. *Biometrika*, 49, 27-40.





## A Methodology for Surveying Disabled Persons using a Supplement to the Labour Force Survey<sup>1</sup>

D. DOLSON, P. GILES, and J.-P. MORIN<sup>2</sup>

### ABSTRACT

In response to a need for data on disabled persons in Canada, Statistics Canada undertook a program to create a disability database. This includes using supplements to the Canadian Labour Force Survey in the Fall of 1983 and the Spring of 1984, as well as including questions on the 1986 Census of Population. A general discussion of the background and content of the survey is presented. A comparison of screening methodologies conducted by Statistics Canada in November 1982 and January 1983 is presented and the results are compared.

**KEY WORDS:** Disability; Screening; Activities of Daily Living.

### 1. INTRODUCTION

On May 23, 1980 the Canadian government created the Special Parliamentary Committee on the Disabled and the Handicapped. In February 1981 (the International Year of the Disabled) this committee published its report, entitled "Obstacles" [4]. Recommendation 113 of the "Obstacles" report reads in part:

"That the Federal Government directs Statistics Canada to give a high priority to the development and implementation of a long-term strategy which will generate comprehensive data on disabled persons in Canada, using population-based surveys and program data."

The government, wishing to respond positively to the recommendations contained in the report, thus requested Statistics Canada to undertake a survey of disabled persons.

This paper focuses on disability surveys conducted as supplements to the Canadian Labour Force Survey (LFS) in October 1983 and June 1984 and on tests which were done in November 1982 and January 1983.

### 2. DEFINITIONS

Definitions developed by the World Health Organization (W.H.O.), given in McWhinnie (1980), were employed by the Special Parliamentary Committee. These definitions arise out of a model which focuses on the consequence of disease, and addresses the following illness-related phenomena.



This paper is a combined version of the two papers entitled "A Methodology for Surveying Disabled Persons Using a Supplement to the Canadian Labour Force Survey" by P. Giles and D. Dolson, and "The Canadian Experience with Screening for Disabled Persons in a Household Survey" by P. Giles, D. Dolson and J.-P. Morin. These papers were presented at the 1983 ASA meetings in Toronto.

P. Giles, Business Survey Methods Division, D. Dolson and J.-P. Morin Institutional & Agriculture Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.



As given in World Health Organization (1980), the definitions of these terms are as follows.

**Impairment:** In the context of the health experience, it is any loss or abnormality of psychological, physiological or anatomical structure or function.

It is characterized by losses or abnormalities that may be temporary or permanent, and that include the existence of an anomaly, defect, or a loss in a limb, organ, tissue, or other structure of the body, including the systems of mental function. Impairment represents the exteriorization of a pathological state, and in principle reflects disturbances at the level of the organ.

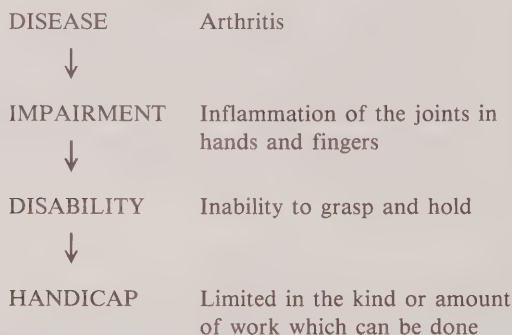
**Disability:** It is any restriction or lack of ability (resulting from an impairment) to perform an activity in the manner or in the range considered normal for a human being.

It is characterized by excesses or deficiencies of customarily expected activity, and may be temporary or permanent, reversible or irreversible, and progressive or regressive. Disabilities may arise as a direct consequence of impairment or as a response by the individual, particularly psychologically, to a physical, sensory, or other impairment. Disability represents the objectification of an impairment, and as such, it reflects disturbances at the level of the person.

**Handicap:** It is a disadvantage for a given individual, resulting from an impairment or disability, that limits or prevents the fulfillment of a role that is normal (depending on age, sex, and social and cultural factors) for that individual.

Handicap is concerned with the value attached to an individual's situation or experience when it departs from the norm. It is characterized by a discordance between the individual's performance or status and the expectations of the individual himself or the particular group of which he is a member. Handicap thus represents the socialization of the impairment or disability, and as such reflects the consequences for the individual (cultural, social, economic and environmental) that stem from the presence of impairment and disability.

To explain these definitions more clearly, consider an example:



### 3. TARGET POPULATION

Ideally the target population would include all persons in Canada who are disabled according to the above definition and subject to constraints on severity and duration. Severity can be regarded in terms of the person; i.e., how severely is a person disabled, or in terms of disability; i.e., how severe is the specific disability. In defining a target population, a measurement of severity must be included, if only implicitly. The duration of disability must be explicitly addressed. To capture all disabilities, including those arising out of acute illnesses of relatively limited duration, would identify a large percentage of people and run contrary to the spirit of the "Obstacles" recommendation. Nevertheless, to limit the population to the permanently disabled is avoiding the issue and ignoring the needs of the long-term but not chronically disabled.

As explained in the next section, the data were collected through the use of the supplementary capacity of the LFS. In addition to the normal constraints of the LFS, one significant limitation was imposed by the use of the LFS. The target population of the disability survey was not to include the "mentally ill". For example, the target population excluded illness such as amnesia, neuroses and phobias but included impairments of intelligence such as mental retardation or dyslexia. It was felt that asking for this information could be very sensitive in nature and negative reactions could compromise the primary objectives of the LFS.

Thus the target population for the disability survey tests includes all persons having one or more physical (nonbehavioural) disabilities, or knowledge acquisition or other educational disabilities (arising from impairments in intelligence, attention, psychomotor functions and language), whose duration has been or is expected to be at least six months. It also includes individuals suffering from diseases of a chronic and degenerative nature and which have a high probability of producing impairments which are physically disabling. In addition the normal constraints of the LFS are in effect which precludes individuals in institutions.

#### 4. DATA COLLECTION

The difficulty is in translating the definitions into a set of questions which identify persons of interest from a set of persons in the general population. This leads to setting an objective to collect information on those who have a high probability of being disabled by any user's definition, and at the same time, keeping the number of people surveyed within reasonable limits.

The first option studied was to include questions on the Census of Population. However the diverse data requirements would have required ten to thirty additional questions, which was clearly not possible. The second option considered was to include a limited number of questions on the Census of Population, which would identify the disabled. In order to meet the data requirements, a follow-up survey would be required. In fact, this option has been approved and disability questions will be included on the 1986 Census questionnaire. This will provide detailed estimates for small areas. However, results from the follow-up survey will not be available before 1988 or 1989. Given that the current demand for data was high as well as the fact that this demand was mainly for national baseline estimates, the Census/follow-up option was considered inadequate by itself.

For meeting current data requirements, two alternatives were considered. The first possibility would be to mount on a continuous or periodic basis a household survey similar to the Canada Health Survey, to provide a profile of the disabled and handicapped. Such a survey could use survey methods designed particularly for the collection of disability data. However, resource constraints made this proposal not feasible. The second alternative was to use the supplementary capacity of the LFS. For reasons of expediency and cost, this was the method chosen. This was the secondary advantage that the resulting disability data on the individual can be directly linked to their labour force data collected by the LFS.

The data collection for the disability survey was conducted in two stages. First, all persons in all households in the LFS sample, except the one-sixth of the sample which is in its first month of the survey, underwent a "screening" process. Persons of potential interest were identified by means of a "screening" questionnaire. The mode of data collection was the same as for the LFS interview. However, the interviewers were asked to obtain non-proxy interviews as often as possible, even if it meant calling back at a later time. This "screened in" population was then asked another set of questions in a follow-up survey. All of these interviews took place about a week after the screening interview. They were all personal interviews and non-proxy responses only were accepted. This second set of questions was designed to collect the data identified as being desirable by a consultation process with the users.

The schedule for the survey was as follows. Three proposed screening questionnaires were tested in November 1982 and January 1983. More details on these surveys will be given later. Based on the results of these surveys, one screening questionnaire was developed. Two "full" surveys with a screen and a more detailed questionnaire as described above were conducted in October 1983 and in June 1984.

## 5. APPROACHES TO SCREENING - OTHER SURVEYS

The first step in constructing a set of screening questions was to investigate experiences encountered by other groups that had previously conducted disability surveys.

The one approach that has been used in many surveys is the Activities of Daily Living (ADL) approach. The Activities of Daily Living are a set of activities which any person is required to perform during the course of their regular living pattern. Although there is no generally recognized "best" set of activities that should be used, the set developed in 1978 by the Organization for Economic and Co-operative Development (OECD) and noted in McWhinnie (1980) has been used by surveys in several countries; see Klaukka (1981), Mizrahi and Mizrahi (1981), Raymond, Christie and Clemence (1981), Van Sonsbeek (1981), Wilson and McNeil (1981).

Since a person's ability to perform an ADL may depend on their use of a physical aid, such as an artificial limb, the use of a list of physical aids for screening could be appropriate.

Another approach for screening is that of major activity limitation. If a person is limited in his/her major activity (i.e., work, school, home) that person is probably experiencing some disability. This approach has been used in the United States in a pretest for a disability survey (1980) and in the annual Health Interview Survey, and in Canada in the Canada Health Survey (1978-79).

A list of chronic conditions could be useful for screening since persons with chronic conditions are in the target population but may be missed by ADL's or activity limitation if the person has intermittent difficulty.

Finally a person could be asked a single self-perception question such as "Do you have any physical disabilities or handicaps?"

## 6. TEST OF SCREENING MECHANISMS

The three Statistics Canada screening tests used combinations of these approaches. Also persons aged 15 or over were administered a different questionnaire than those under 15 years of age. No suitable set of ADL's has been compiled for children. In fact, most disability surveys that have been previously conducted have excluded children. Here, we will consider only persons aged 15 years and older.

In the November 1983 Labour Force Survey, each respondent was asked "Does ... now have any disability or handicap which has lasted or is expected to last six months or more?" This was called Test 1. Persons screened in by this question were those responding "yes".

The other approaches to screening were tested using two different questionnaires each administered in the January 1983 Labour Force Survey. These questionnaires were called Test 2 and Test 3.

Test 2 included the following sections: a list of special aids, a list of ADL's, and the activity limitation question "Are there any (other) conditions or health problems that now prevent or limit ... when carrying out his/her normal daily activities at a job, in school, or in the home? Please report only difficulties which are expected to last more than six months". Persons who reported using at least one of the special aids or having trouble doing at least one of the ADL's or who answered "yes" to the above question were screened in.



Test 3 included a list of ADL's, a list of chronic conditions, and the following two work disability questions: "Is . . . limited in the kind or amount of work he/she can do at his/her job or business because of a long-term physical condition or health problem?" (asked only to employed persons) and "Is . . . prevented or limited in the kind or amount of work he/she could do at any job or business because of a long-term physical condition or health problem?". Persons who reported having trouble doing at least one of the ADL's or who had at least one of the chronic conditions or who replied "yes" to either of the above two questions were screened in.

Lists of the special aids, the chronic conditions, and the ADL's used in these tests are given in the appendix. It should be noted that the ADL's used in Test 2 and Test 3 were identical although slightly modified from the OECD list. In addition, each test takes a different approach to the ADL list. Test 2 permits the use of aids to perform the activities, while Test 3 does not.

These proposed screening methods do not permit an assessment of whether or not the target population is being correctly identified, unless they are used on a control population. This has not been done for the Canadian survey.

Test 1 was administered to all persons in households in the November 1982 Labour Force Survey. In January 1983, two rotation groups were used for each of Test 2 and Test 3. These rotation groups were chosen so that each in-sample household had also been in-sample in November 1982. This facilitated the comparison of the results of Test 1 with those of Test 2 and Test 3.

## 6.1 Major Findings of the Pilot Tests

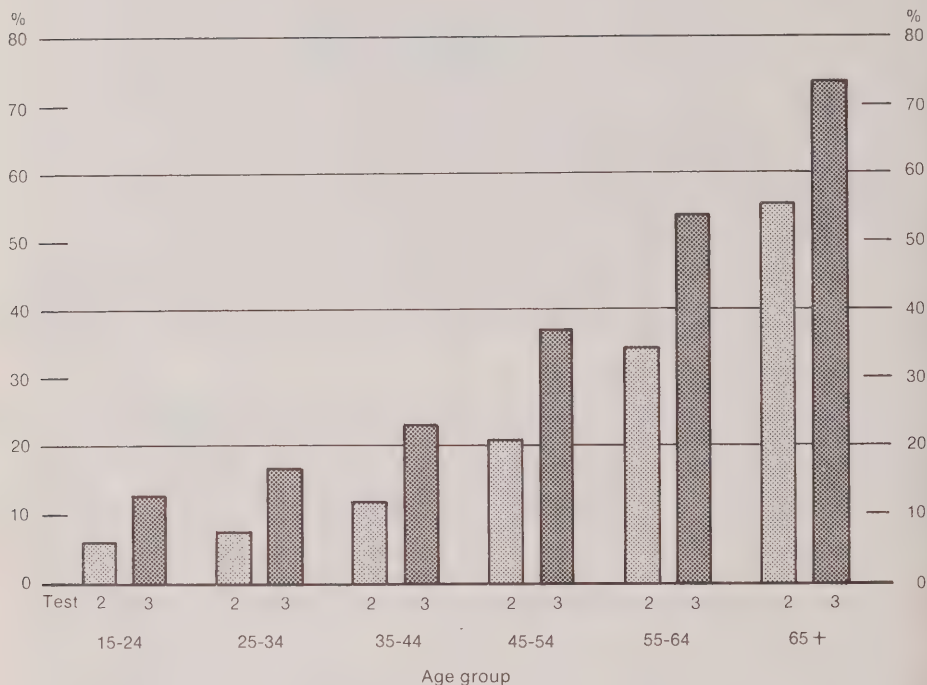
The major goal in the analysis of the tests was to determine the set of questions that would be most effective in screening in those persons who belong to the target population. Another important factor was to determine an effective screen that would also not unduly increase respondent burden or cost. In addition, sources of non-sampling errors that became evident during the analysis were noted so that, for example, survey procedures or questionnaire design could be changed appropriately. The following discussion presents the major findings of the analysis.

The sample size for Test 1 was about 115,000 persons. For each of Test 2 and Test 3 the sample size was almost 38,000 persons. For all three tests the samples were about 49% male and 51% female.

Table 1 shows by sex the percentage of the sample screened in by Test 1, Test 2, Test 3 and by each section of Test 2 and Test 3. Most notable in this table is that Test 1 screened in only 5.6% of the sample as compared to about 16% of the sample for either of the ADL questions. Given that functional limitation as measured by the activities of daily living is a key indicator of disability, this shows that the single question asked in Test 1 is not effective for screening in the entire target population.

**Table 1**  
Percent Screened in by Each Section of  
Each Questionnaire, by Sex

Section	Percent Screened in	
	Male	Female
Test 1: . . . . .	5.5	5.7
Test 2: ADL . . . . .	14.7	16.2
Test 2: Aids . . . . .	3.2	2.9
Test 2: Activity Limitation . . . . .	5.9	6.1
Test 2: . . . . .	18.3	19.4
Test 3: ADL . . . . .	15.4	16.7
Test 3: Work Disability . . . . .	13.0	13.1
Test 3: Chronic Conditions . . . . .	25.6	27.4
Test 3: . . . . .	29.8	31.2



**Figure 1.** Percent Screened in by Test 2 and Test 3

Percentages of the sample screened in by Test 2 and by Test 3 are shown by age in Figure 1. For both tests, the probability of being screened in is highly related to age. The probability of being screened in is an increasing function of age. Although the degree varies, the same relationship holds for each of the Sections of Test 2 and Test 3, for each of the seventeen ADL's and for most of the chronic conditions (multiple sclerosis, epilepsy, cerebral palsy, cystic fibrosis and muscular dystrophy are exceptions).

Since functional limitation is a good indicator of disability the question for the analysis was how to include the ADL's on the screen rather than whether to include them. The "with special aid approach" (Test 2) selected 15.5% of the sample while the "without special aid approach" (Test 3) selected 16.0% of the sample. The difference between these figures is not statistically significant. For both respondents and interviewers the "with special aid" concept seemed to be more natural and more easily understood and was therefore chosen for the finalized screen.

Since the complete set of activities can be used to obtain a measure of degree of disability, all seventeen ADL's were retained for the finalized screen.

The special aids section in Test 2 screened in 3.0% of the sample. Of these, 84.2% were screened in by the ADL section. Thus although this type of data is of interest for the disability database, the aids section is not an efficient screening mechanism, especially in combination with the ADL section. Consequently, questions on aids are not included on the finalized screen.

In addition to activities of daily living, major activity limitation is also an important aspect of disability. The Test 3 questionnaire addressed this by the two questions noted earlier in this section. These questions, however, considered it only from the point of view of work disability. Of the persons in the Test 3 sample, 13.1% were screened in by these questions (6.8% of employed persons, 6.6% of unemployed persons and 22.7% of persons not in the labour force).

Although there were no obvious problems with the data from the work disability questions, there were some operational difficulties. In particular, the questions sometimes seemed irrelevant to retired persons. Thus, the finalized major activity limitation question was adapted in order to better suit persons not in the labour force.

The chronic conditions section in Test 3 screened in 23.9% of the Test 3 sample. Of these, 37.9% were not otherwise screened in. There were two main problems with this section. First, the question was difficult to answer for respondents who were not sure of the nature of their particular condition(s). Another difficulty is that the data sought on the follow-up questionnaire, are generally more pertinent to persons who are currently disabled. Thus persons screened in by a chronic condition, but not currently having a functional limitation or a major activity limitation would be interviewed for the follow-up and probably provide little useful data.

Given these problems, chronic conditions are not used as screening criteria on the finalized screen. One exception to this is mental handicap. This one condition is retained as a screening item since there may be persons with mental handicaps who are not screened in by the ADL's, or even by the major activity limitation question.

The project team felt that there would likely be differences in proxy and non-proxy responses related to any particular person. To increase non-proxy response, interviewers were instructed that whenever possible the questionnaire was to be completed by interviewing the individual to whom it applied. If a knowledgeable household member insisted upon responding for other household members, then this response was to be accepted; although the practice was to be avoided.

The level of proxy response obtained is considered to be fairly low (20.7% for females, 32.2% for males). Even after accounting for age-sex differences, it was found that proxy respondents were slightly less likely to be screened in than non-proxy respondents. Two reasons can be suggested as to why the probabilities of selection differ. First, persons who are unavailable and for whom proxy responses were provided may be less likely to be disabled. Second, proxy respondents may be less likely to state that a person has trouble doing an ADL or a major activity than the person himself/herself.

7. DATA REQUIREMENTS

As a result of a solicitation of data requirements from users, 173 responses were received which identified 588 issues of data needs. The following eleven areas were identified.

Issue	Number of users requesting data
1. Nature of impairment	123
2. Demographic characteristics	95
3. Employment	85
4. Assistance	77
5. Education	50
6. Accommodation	45
7. Economic Characteristics	41
8. Transportation	29
9. Social activities	26
10. Health	9
11. Communication	8



The nature of impairment/disability/handicap is basic to the survey. Considerable detail is collected, including cause of disability. Most users of the data are interested in focusing on the impairment or disability groupings separately. Demographic characteristics are always important data as they allow the user to identify sectors of the population falling into different categories.

It can be easily understood that employment data about the disabled would be an important issue, as employment is a key component to the independent living of a disabled person. A great deal of employment data are already collected by the LFS. The follow-up survey will collect data related to employment limitations experienced as a result of the disability. In addition to the analysis of the data for the disabled population, these data will permit comparisons of the labour market characteristics of the disabled with those of the Canadian population as a whole.

Three aspects of assistance are considered: technical aids and skills, employment related assistance, and education related assistance. In all three areas, need for aids or assistance was deemed more important than was use. Under technical aids and special skills, interest is greatest for those aids and skills which are most prevalent, or for which special services or facilities must be provided. The aids would be grouped under hearing, speaking, seeing and mobility. Employment related assistance refers to the impact of aids on the ability of the disabled to work.

The LFS already determines the highest level of education achieved by each respondent. In addition, the follow-up survey will collect data on current educational activity and the impact of disability on current and past education.

The LFS collects information on the dwellings of the respondents. Additional accommodation data will be collected on special architectural/structural features, both inside and outside the home and other buildings.

Economic characteristics will be considered in the following areas: personal income including financial assistance received due to disability, sources of financial assistance, and special expenses incurred as a result of the disability.

Transportation data will be collected on three types of travel: travel to work or school, other local travel and long distance travel. Details on each area will identify the modes of transportation used, frequency of use and problems encountered due to the disability.

Although some interest was expressed by users in data on social/leisure activities, health and communication, no data will be collected for these issues by the present survey. For the first and third of these issues it was felt that reliable and useful data could not be collected in this survey. Questions related to health are also not included because of the already substantial response burden imposed by issues of higher priority.

## **8. RELIABILITY OF ESTIMATES FROM THE DISABILITY SURVEY**

When determining the content of a questionnaire, consideration must be given to the reliability of estimates produced for the various data items. It is useless to collect data which will not be reliable enough to publish, even if the data requirement has a high priority. The reliability of an estimate is tied directly to the sample size. For this survey, the number of persons receiving a screening questionnaire is fixed. Therefore the reliability of the estimates produced will depend on the number of disabled falling into the sample and the prevalence rates of each characteristic of interest. Based on population projections from the 1981 Census of Population and certain assumptions it is possible to estimate minimum prevalence rates required to produce an estimate which is "reliable enough" to publish. An estimate whose coefficient of variation is less than or equal to 16.5% is considered releasable without qualification by LFS. Table 2 displays the expected minimum releasable estimates for the disability survey. Estimates of this size or higher will have coefficients of variation of less than 16.5%, subject to the validity of the following assumptions.

- (1) All LFS sampled households are administered the screening questionnaire except the one-sixth of the sample which is in its first month of the survey,
- (2) 2.95 persons per household on average,
- (3) 5% LFS non-response rate,
- (4) 5% disability survey non-response rate,
- (5) Design effect of 2.5 (this accounts for the fact that a simple random sample design was not used),
- (6) 19% of total adult population and 8% of total child population (aged less than 15) are screened in.

To explain the table in more detail, consider, for example the province of Newfoundland. An estimate of 9,000 persons possessing a particular characteristic will have a coefficient of variation less than 16.5% and is publishable whereas an estimate of 7,000 will have a coefficient of variation greater than 16.5% and is not publishable. An estimate of 8,000 is approximately 1.4% of the population of Newfoundland. Given the assumptions about percentage disabled in the population, an estimate of 8,000 is approximately 10.1% of the adult disabled population and 59.1% of the child disabled population of Newfoundland.

The design effects observed from Test 2, Test 3 and the October 1983 Disability Survey for number of persons screened in were about 1.5. This suggests that design effects for number of screened in persons with specified characteristics are probably also much less than 2.5.

In the October 1983 Disability Survey 12.9% of adults and 4.8% of children in the sample were screened in.

Table 2  
The expected minimum releaseable estimates

Province/Region	Min P <sup>a</sup>	Min X <sup>b</sup>	Min D <sup>c</sup>	
			Adults	Children
Atlantic .....	0.4	7,500	2.3	16.2
NFLD .....	1.4	8,000	10.1	59.1
PEI .....	2.9	3,500	20.0	>100
NS .....	1.1	8,500	6.9	54.5
NB .....	1.0	7,000	6.7	49.5
Quebec .....	0.5	30,500	3.3	28.2
Ontario .....	0.4	32,500	2.6	22.0
Prairies .....	0.3	10,000	1.7	12.4
MAN .....	0.9	9,000	7.0	47.6
SASK .....	0.8	7,000	5.0	38.0
ALTA .....	0.6	13,000	4.1	29.9
British Columbia .....	0.7	17,500	4.4	38.2
Canada .....	0.1	18,000	0.6	4.2

<sup>a</sup> Min P = minimum estimable percentage of the total population,  
<sup>b</sup> Min X = minimum estimable total,  
<sup>c</sup> Min D = minimum estimable percentage of disabled adults or children.

## APPENDIX

### Special Aids

- Does . . . now use
- a wheelchair?
  - crutches or other walking aids?
  - any kind of brace excluding braces for teeth?
  - medically prescribed orthopedic shoes?
  - artificial limb(s)?
  - a hearing aid?
  - a guide dog?
  - a white cane?
  - any other kind of special aid?

### Activities of Daily Living

- Does . . . now have any trouble
- walking 400 metres without resting (about 3 city blocks)?
  - walking up and down a flight of stairs?
  - carrying an object of 5 kg. 10 metres (e.g. carrying a 12 lb. bag of groceries 30 ft.)?
  - moving from one room to another?
  - standing for long periods of time (e.g. more than 20 minutes)?
  - when standing, bending down and picking up an object from the floor (e.g. a shoe)?
  - dressing and undressing himself/herself?
  - getting in and out of bed?
  - cutting own toenails?
  - using fingers to grasp or handle?
  - reading?
  - cutting own food?
  - reading ordinary newsprint (with glasses if normally worn)?
  - seeing clearly the face of someone from 4 metres (e.g. across a room) (with glasses if normally worn)?
  - hearing what is said in a normal conversation with one other person?
  - hearing what is said in a normal conversation with at least two other persons?
  - speaking and being understood?

### Chronic Conditions

- Which, if any, of these long term conditions or health problems does . . . presently have?
- heart disease
  - kidney disease
  - lung disease
  - cancer
  - diabetes
  - epilepsy
  - cerebral palsy
  - multiple sclerosis
  - cystic fibrosis
  - muscular dystrophy
  - paralysis of any kind
  - arthritis or rheumatism of a serious nature
  - high blood pressure
  - hearing trouble (uncorrected by aid)
  - vision trouble (uncorrected by aid)
  - mental handicap
  - any missing limb(s) including finger(s) and toe(s)
  - any other long-term condition or health problem (please specify)

## REFERENCES

- STATISTICS CANADA. (1983). Data Content for the Statistics Canada Survey of the Disabled. *Disability Database Development Project*, Health Division, technical report.
- CARTER, R.G., GILES, P.D., and SHERIDAN, M.J. (1982). Description and Rationale for the Screen Tests for the January 1983 Disability Survey. *Disability Database Development Project*, Health Division, Statistics Canada.
- GRABOWIECKI, F. (1982). Discussion of the Target Population for the Disability Survey. *Disability Database Development Project*, Health Division, Statistics Canada.
- HOUSE OF COMMONS. (1981). Obstacles. *Report of the Special Committee on the Disabled and Handicapped*. Ottawa.
- MORIN, J.-P. (1983). Enquête sur les handicapés compte rendu des requêtes des utilisateurs. *Disability Database Development Project*, Health Division, Statistics Canada.
- WORLD HEALTH ORGANIZATION. (1980). International Classification of Impairments, Disabilities and Handicaps. Geneva, Switzerland.
- McWHINNIE, J.R. (1980). Disability Indicators for Measuring Well-being. *OECD Social Indicators Programme Technical Report Series*, Paris, France.
- McDOWELL, I. (1981). An Examination of the OECD Survey Questions in a Canadian Study. *Revue d'épidémiologie et de santé publique*, 29, 421-429.
- KLAUKKA, T. (1981). Application of the OECD Disability Questions in Finland. *Revue d'épidémiologie et de santé publique*, 29, 431-439.
- MIZRAHI, A. and MIZRAHI, A. (1981). Évaluation de l'état de santé de personnes âgées en France, à l'aide de plusieurs indicateurs, dont les questions de l'OCDE. *Revue d'épidémiologie et de santé publique*, 29, 441-450.
- RAYMOND, L., CHRISTE, E., CLEMENCE, A. (1981). Vers l'établissement d'un score global d'incapacité fonctionnelle sur la base des questions de l'OCDE, d'après une enquête en Suisse. *Revue d'épidémiologie et de santé publique*, 29, 451-459.
- VAN SONSBECK, J.L.A. (1981). Applications aux Pays-Bas des questions de l'OCDE relatives à l'incapacité. *Revue d'épidémiologie et de santé publique*, 29, 461-468.
- McWHINNIE, J.R. (1981). Disability Assessment in Population Surveys: Results of the OECD Common Development Effort. *Revue d'épidémiologie et de santé publique*, 29, 413-419.
- WILSON, R.W., McNEIL, J.M. (1981). Preliminary Analysis of OECD Disability on the Pretest of the post Census Disability Survey. *Revue d'épidémiologie et de santé publique*, 29, 469-475.



'Logistic Regression Analysis of Labour Force Survey Data' by S. Kumar and J.N.K. Rao, *Survey Methodology* (1984), 10, 62-81.

In the formula (26) for the projection matrix  $M = I - H$ , the matrix  $\hat{V}_i$  should have been given as

$$\text{diag } [\hat{p}_1(1-\hat{p}_1)(nw_1), \dots, \hat{p}_J(1-\hat{p}_J)(nw_J)]$$

instead of  $\text{diag } [\hat{p}_1(1-\hat{p}_1)/(nw_1), \dots, \hat{p}_J(1-\hat{p}_J)/(nw_J)]$ . Because of this error, the figures 3 and 4, based on the diagonal elements  $m_{ii}$  and  $h_{ii}$  of  $M$  and  $H$  respectively, are incorrect. The corrected figures are given here as figures 3\* and 4\* which indicate that cells numbered 2, 3 and 55 warrant further examination. It may be noted that the diagnostics for assessing the impact of extreme points on the fit (see figures 5-10, pp. 79-81) have also identified cells 2 and 3 as extreme points, excepting figure 8. The overall observation in the paper that cells 2, 3 and 7 may be possible candidates for deletion is unaffected by the above error.

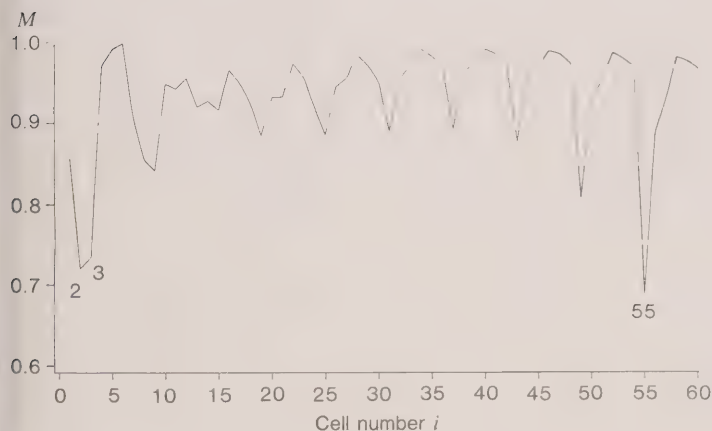


Figure 3\*. Index plot of  $m_{ii}$ .

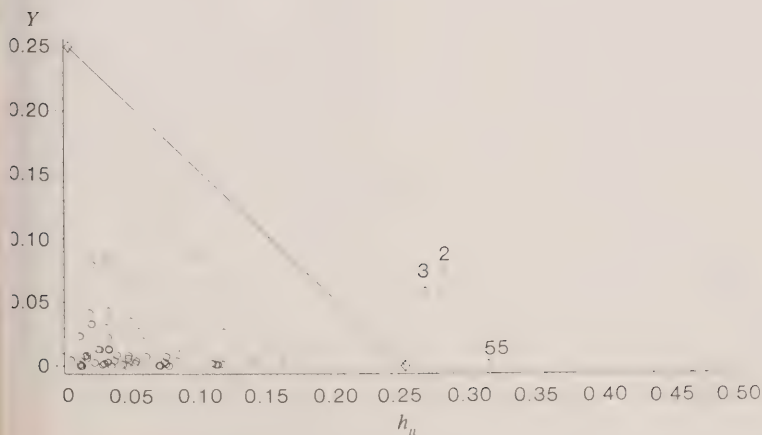


Figure 4\*. Scatter plot of  $\chi^2_i/\chi^2$  vs.  $h_{ii}$ .





## ACKNOWLEDGEMENTS

The Survey Methodology Journal wishes to thank P. Pariseau and C. Paslawski for their patient preparation of the manuscripts and to D. Lemire for his service as liaison with the translation and the production services.

Acknowledgement is also due to H. Gough, I. Trottier, M. Brodeur and G. Lemaître for their help in proofreading of French version of papers.

Finally, the Journal wishes to thank the following persons who have served as referees during the past year.

M. Bankier

M. Lawes

G. Brackstone

D. Paton

N. Chinnappa

C. Patrick

D. Drew

D. Rhoades

J. Gambino

K.P. Srinath

J. Kovar

R. Verma



## GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of *Survey Methodology* as a guide and note particularly the following points:

### 1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ( $8\frac{1}{2} \times 11$  inch), one side only, entirely double spaced with margins of at least  $1\frac{1}{2}$  inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

### 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

### 3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(-)" and "log(-)", etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w,  $\omega$ ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

### 4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

### 5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

## REMERCIEMENTS

La revue *Techniques d'enquête* désire remercier P. Parisseau et C. Paslawski qui ont soigneusement préparé la revue, et D. Lemire qui a assuré une liaison constante entre les services de traduction et les services de production.

Nous remercions aussi H. Gough, I. Trotter, M. Brodeur et G. Lemaître pour leur contribution lors de la révision de la traduction des textes.

Finalement, la revue désire remercier les personnes suivantes qui ont accepté de faire la critique des articles présentés au cours de l'année dernière.

M. Bankier	M. Lawes
G. Brackstone	D. Paton
N. Chinnappa	C. Patrick
D. Drew	D. Rhoades
J. Gambino	K.P. Srinath
J. Kovar	R. Verma





'Régression logistique et analyse de données de l'enquête sur la population active' par S. Kumar et J.N.K. Rao. Techniques d'enquête (1984), 10, 68-90.

Dans la formule (26) de a matrice-projection  $M = I - H$ , la matrice  $V_b$  aurait dû être exprimée:

$$\text{diag} [b'_1(1 - \beta_1)/(nw_1), \dots, b'_j(1 - \beta_j)/(nw_j)]$$

au lieu de  $\text{diag} [b'_1(1 - \beta_1)/(nw_1), \dots, b'_j(1 - \beta_j)/(nw_j)]$ . A cause de cette erreur, les figures 3 et 4, construites à l'aide des éléments diagonaux  $m''_j$  et  $h''_j$  de  $M$  et  $H$  respectivement, sont inexactes. Il faut plutôt se reporter aux figures 3\* et 4\* qui révèlent que les cas numéros 2, 3 et 55 exigent un examen supplémentaire. Il convient de souligner que les indices diagnostiques servant à évaluer l'incidence des points extrêmes sur la qualité de l'ajustement (voir figures 5-10, p. 87-89) ont permis de détecter également comme extrêmes les cas 2 et 3, sauf dans le cas de la figure 8. L'erreur qui s'est glissée dans la formule (26) ne modifie en rien l'observation générale formulée dans ce document, à savoir qu'il serait peut-être opportun de supprimer les cas 2, 3 et 7.

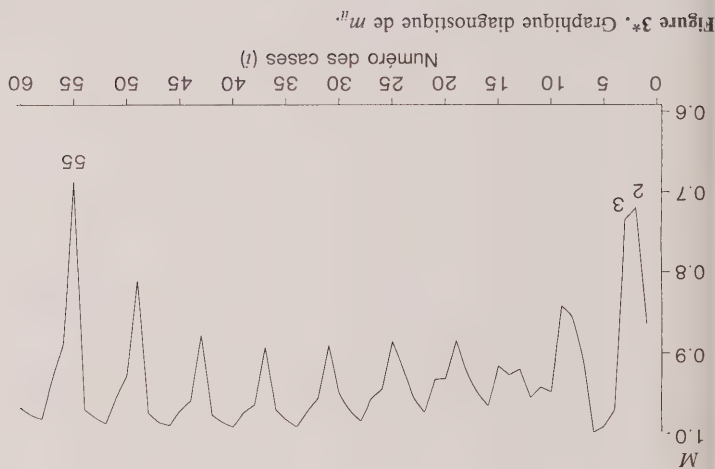


Figure 3\*. Graphique diagnostique de  $m''_j$ .

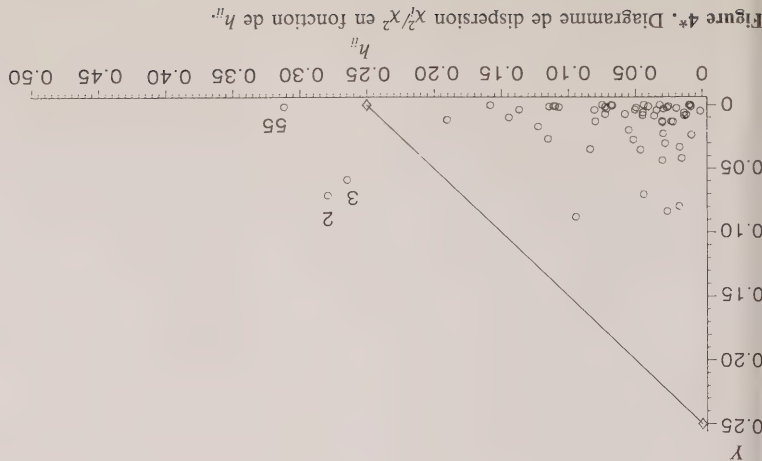


Figure 4\*. Diagramme de dispersion  $X^2_j/X^2$  en fonction de  $h''_j$ .

## BIBLIOGRAPHIE

- STATISTIQUE CANADA. (1983) Contenu des données de l'enquête de Statistique Canada auprès des handicapés. *Projet d'établissement d'une base de données sur les invalides*, Division de la santé rapport technique.
- CARTER, R. G., GILES, P. D., and SHERIDAN, M. J. (1982). Description and Rationale for the Screen Tests for the January 1983 Disability Survey. *Projet d'établissement d'une base de données sur les invalides*, Division de la santé, Statistique Canada.
- GRABOWIECKI, F. (1982). Discussion of the Target Population for the Disability Survey. *Projet d'établissement d'une base de données sur les invalides*, Division de la santé, Statistique Canada.
- CHAMBRE DES COMMUNES. (1981). Obstacles, *Rapport du Comité spécial concernant les invalides et les handicapés*. Ottawa.
- MORIN, J.-P. (1983). Enquête sur les handicapés compte rendu des requêtes des utilisateurs. *Projet d'établissement d'une base de données sur les invalides*, Division de la santé, Statistique Canada.
- ORGANISATION MONDIALE DE LA SANTÉ. (1980). Classification internationale des déficiences, incapacités et handicaps. Genève, Suisse.
- McWHINNIE, J. R. (1980). Disability Indicators for Measuring Well-being. *OCDE Social Indicators Programme Technical Report Series*, Paris, France.
- McDOWELL, I. (1981). Un examen des questions proposées par l'OECD dans le cadre de l'enquête canadienne. *Revue d'épidémiologie et de santé publique*, 29, 421-429.
- KLAVUKKA, T. (1981). Application en Finlande des questions de l'OCDE relatives à l'incapacité. *Revue d'épidémiologie et de santé publique*, 29, 431-439.
- MIZRAHI, A. and MIZRAHI, A. (1981). Evaluation de l'état de santé de personnes âgées en France, à l'aide de plusieurs indicateurs, dont les questions de l'OCDE. *Revue d'épidémiologie et de santé publique*, 29, 441-450.
- RAYMOND, L., CHRISTE, E., CLEMENCE, A. (1981). Vers l'établissement d'un score global d'incapacité fonctionnelle sur la base des questions de l'OCDE, d'après une enquête en Suisse. *Revue d'épidémiologie et de santé publique*, 29, 451-459.
- VAN SONSBEEK, J. L. A. (1981). Applications aux Pays-Bas des questions de l'OCDE relatives à l'incapacité. *Revue d'épidémiologie et de santé publique*, 29, 461-468.
- McWHINNIE, J. R. (1981). Evaluation de l'incapacité dans les enquêtes épidémiologiques: Résultats du projet coopératif de l'OCDE visant à développer cette évaluation. *Revue d'épidémiologie et de santé publique*, 29, 413-419.
- WILSON, R. W., McNEIL, J. M. (1981). Analyse préliminaire de l'incapacité au sein de l'OCDE, d'après le test initial de l'enquête post-censitaire. *Revue d'épidémiologie et de santé publique*, 29, 469-475.

**Appareils spéciaux**

... utilise-t-il (elle) ...

un fauteuil roulant?

des béquilles ou d'autres appareils pour marcher?

un appareil orthopédique, sauf un appareil dentaire?

des chaussures orthopédiques prescrites par un médecin?

(un (des) membre(s) artificiel(s)?)

une prothèse auditive?

un chien guide?

une canne blanche?

une autre genre d'appareil spécial?

**Activités de la vie quotidienne**

... éprouve-t-il (elle) des difficultés à ...

Marcher sur une distance de 400 mètres sans se reposer (environ 3 pâtés de maison)?

Monter et descendre un escalier?

Transporter un objet de 5 kg sur 10 mètres (c'est-à-dire transporter un sac d'épicerie de 12 livres sur une distance de 30 pieds)?

Se déplacer d'une pièce à une autre?

Se tenir debout pendant de longues périodes (par ex., pendant plus de 20 minutes)?

En position debout, se pencher et ramasser un objet à partir du plancher (par ex., un soulier)?

S'habiller et se déshabiller?

Se mettre au lit et sortir du lit?

Se couper les ongles d'orteils?

Se servir de ses doigts pour saisir ou manier un objet?

Lire?

Couper ses aliments?

Lire des caractères ordinaires (avec des verres si elle (il) en porte habituellement)?

Voir clairement la figure de quelqu'un à 4 mètres (par ex., d'un bout à l'autre d'une pièce)

(avec des verres si elle (il) en porte habituellement)?

Entendre ce qui se dit au cours d'une conversation normale avec une autre personne?

Entendre ce qui se dit au cours d'une conversation normale avec au moins deux personnes?

Parler et être compris?

**Affections chroniques**

Parmi les affections chroniques suivantes, quelles sont celles, le cas échéant, dont ... souffre?

Maladie cardiaque

Maladie rénale

Maladie pulmonaire

Cancer

Diabète

Épilepsie

Paralysie cérébrale

Sclérose en plaques

Fibrose kystique

Dystrophie musculaire

Paralysie de tout genre

Arthrite ou rhumatisme aigus

Hypertension

Trouble de l'ouïe (non corrigé par une prothèse)

Trouble de la vue (non corrigé par une prothèse)

Handicap mental

Un (des) membre(s) absent(s), y compris un (des) doigt(s) et un (des) orteils

Autres affections chroniques ou problèmes de santé à long terme (préciser)

Tableau 3

Estimations minimales publiables prévues

Province/région		Min P <sup>a</sup>	Min X <sup>b</sup>	Min D <sup>c</sup>		
					Adultes	Enfants
Provinces de l'Atlantique	.....	0.4	7,500	2.3	16.2	
I.-N.	.....	1.4	8,000	10.1	59.1	
I.-P.-E.	.....	2.9	3,500	20.0	>100	
N.-E.	.....	1.1	8,500	6.9	54.5	
N.-B.	.....	1.0	7,000	6.7	49.5	
Québec	.....	0.5	30,500	3.3	28.2	
Ontario	.....	0.4	32,500	2.6	22.0	
Prairies	.....	0.3	10,000	1.7	12.4	
MAN	.....	0.9	9,000	7.0	47.6	
SASK	.....	0.8	7,000	5.0	38.0	
ALTA	.....	0.6	13,000	4.1	29.9	
Colombie-Britannique	.....	0.7	17,500	4.4	38.2	
Canada	.....	0.1	18,000	0.6	4.2	

<sup>a</sup> Min P = pourcentage minimum estimable de l'ensemble de la population,

<sup>b</sup> Min X = total minimum estimable,

<sup>c</sup> Min D = pourcentage minimum estimable de l'ensemble de la population des adultes ou des enfants handicapés.

2) Chaque ménage compte en moyenne 2.95 personnes.

3) Le taux de non-réponse à l'EPA est de 5%.

4) Le taux de non-réponse à l'enquête sur les personnes handicapées est de 5%.

5) La valeur observée de l'effet du plan est de 2.5 (ce qui tient au fait qu'on n'a pas eu recours à un échantillonnage aléatoire simple).

6) 19% de l'ensemble de la population adulte et 8% de toute la population d'enfants (âgés de moins de 15 ans) sont repérés.

Pour expliquer ce tableau, prenons le cas de la province de Terre-Neuve. Un total estimatif de 9,000 personnes possédant une caractéristique particulière auront un coefficient de variation de moins de 16.5%, ce qui fait que les résultats sont publiables, alors qu'un total estimatif de 7,000 personnes auront un coefficient de variation supérieur à 16.5%, ce qui fait que les résultats ne peuvent être publiés. Un total estimatif de 8,000 personnes représente environ 1.4% de l'ensemble de la population de Terre-Neuve. Selon les hypothèses au sujet du pourcentage de personnes handicapées dans la population, un total estimatif de 8,000 représente environ 10.1% de la population de personnes handicapées adultes et 59.1% de la population d'enfants handicapés de Terre-Neuve.

On a observé des valeurs de l'effet du plan de 1.5 environ dans le cas des personnes déprimées à l'aide du test 2, du test 3 et de l'enquête sur les handicapés d'octobre 1983. Cela laisse supposer que l'effet du plan dans le cas des personnes déprimées et possédant des caractéristiques données sera probablement très inférieur à 2.5.

L'enquête sur les handicapés menée en octobre 1983 a permis de repérer 12.9% des adultes et 4.8% des enfants de l'échantillon.

Trois aspects de l'aide sont envisagés : les aides et aptitudes techniques, l'aide liée à un emploi et l'aide liée à l'éducation. Dans ces trois secteurs, les besoins ont été jugés plus importants que l'usage proprement dit. Dans le cas des aides techniques et des aptitudes spéciales, ce sont surtout les aides et aptitudes les plus courantes ou qui nécessitent des installations spéciales ou des services particuliers qui retiennent l'attention. Les aides sont regroupées selon qu'il s'agit d'appareils pour l'audition, l'élocution, la vue et la mobilité. L'aide liée à un emploi renvoie à l'incidence des appareils et aides sur la capacité de la personne souffrant d'une incapacité de travailler.

L'EPA détermine déjà le niveau d'instruction de chaque répondant. En outre, l'enquête de suivi permettra d'obtenir des données sur les études actuelles et l'effet de l'incapacité sur les études actuelles et antérieures.

L'EPA recueille des données sur les conditions de logement des répondants. D'autres éléments d'information sur la question porteront sur les caractéristiques architecturales/structurales, à l'extérieur comme à l'intérieur de la maison et des autres bâtiments. Les renseignements d'ordre économique auront trait aux questions suivantes : revenu personnel, y compris l'aide financière reçue en raison de l'incapacité, les sources d'aide financière et les dépenses entraînées par l'incapacité.

Les données sur le transport seront recueillies sur trois types de déplacements : les déplacements vers le lieu de travail ou l'établissement d'enseignement, les autres déplacements locaux et les longs parcours. À l'intérieur de chaque domaine, on déterminera les modes de transport utilisés, la fréquence d'utilisation et les problèmes découlant de l'incapacité.

Bien que les utilisateurs aient montré un certain intérêt pour des éléments d'information sur les activités récréatives/sociales, la santé et les communications, la présente enquête n'aura pas pour objet de recueillir des données sur ces questions. Pour la première et la troisième de ces questions, il a été jugé qu'aucune donnée fiable et utile ne pouvait être obtenue dans le cadre de cette enquête. Les questions liées à la santé ne sont pas plus retenues étant donné le fardeau déjà lourd imposé aux répondants par les questions puis prioritaires.

## 8. FIABILITÉ DES ESTIMATIONS FONDÉES SUR L'ENQUÊTE SUR LES PERSONNES HANDICAPÉES

Lorsqu'on détermine le contenu d'un questionnaire, on doit se préoccuper de la fiabilité des estimations qui seront produites relativement aux divers éléments d'information. Il est en effet inutile de recueillir des données qui ne seront pas suffisamment fiables pour être publiées, même si l'obtention de ces données constitue une grande priorité. La fiabilité d'une estimation est directement rattachée à la taille de l'échantillon. Pour les besoins de cette enquête, le nombre de personnes à qui un questionnaire de sélection est adressé est déterminé d'avance. Voilà pourquoi la fiabilité des estimations dépendra du nombre de personnes handicapées faisant partie de l'échantillon et des taux de prévalence, pour chaque caractéristique intéressant les utilisateurs. En se fondant sur les projections démographiques du recensement de 1981 et sur certaines hypothèses, il est possible d'établir approximativement les taux de prévalence minimaux nécessaires pour produire une estimation qui soit "suffisamment fiable" pour être publiée. Une estimation dont le coefficient de variation est inférieur ou égal à 16,5% est jugée publiable sans réserves. Le tableau reproduit ci-dessous fait état des estimations minimales publiables qu'on prévoit obtenir dans le cas de l'enquête sur les personnes souffrant d'une incapacité. Les estimations de cette importance ou supérieures à celle-ci auront un coefficient de variation inférieur à 16,5%, sous réserve de la validité des hypothèses énoncées ci-dessous.

(1) Tous les ménages qui font partie de l'échantillon de l'EPA doivent répondre au questionnaire de sélection, exception faite du sixième de l'échantillon pour qui il s'agit du premier mois de l'enquête.



Le groupe de travail a jugé qu'il y aurait vraisemblablement des différences entre les réponses données par l'intéressé et celles qui sont données par quelqu'un qui répond en son nom. Pour augmenter le nombre des réponses données par l'intéressé, on a demandé aux interveners de veiller, dans la mesure du possible, à remplir le questionnaire en interviewant la personne concernée. Si un membre bien informé du ménage insistait pour répondre à la place de autres membres du ménage, alors ses réponses devaient être acceptées, bien que cette pratique devrait être évitée.

Le taux de réponse provenant d'enquêtés-substitués est considéré assez faible (20,7% chez les femmes et 32,2% chez les hommes). Même après avoir tenu compte des différences selon l'âge et le sexe, on a trouvé que les enquêtés dont les réponses ont été obtenues par un substitué étaient légèrement moins susceptibles d'être dépistés que les enquêtés répondant eux-mêmes. On peut énoncer deux raisons pour lesquelles les probabilités d'être dépistés diffèrent selon qu'il la personne répond elle-même ou par personne interposée. Premièrement, les personnes qui ne sont pas accessibles et pour lesquelles des réponses ont été obtenues par personne interposée ont peut-être moins de probabilités d'être handicapées. Deuxièmement, les personnes substitués sont peut-être moins enclins que la personne intéressée à déclarer qu'une personne éprouve de la difficulté à accomplir une AVQ ou une activité principale.

7. BESOINS EN DONNÉES

On a demandé aux utilisateurs des données de faire connaître leurs exigences en matière de données et on a obtenu 173 réponses qui faisaient état de 588 secteurs où les besoins de données se faisaient sentir. On a ainsi pu définir onze grands domaines.

Domaine	Nombre d'utilisateurs réclamant des données
1. Nature de la déficience	123
2. Caractéristiques démographiques	95
3. Emploi	85
4. Aide	77
5. Education	50
6. Logement	45
7. Caractéristiques économiques	41
8. Transport	29
9. Activités sociales	26
10. Santé	9
11. Communications	8

La nature de la déficience, de l'incapacité ou du handicap revêt une importance primordiale du point de vue de l'enquête. Les enquêteurs cherchent à obtenir sur cette question énormément de détails, notamment sur la cause de l'incapacité. La plupart des utilisateurs des données veut considérer séparément les catégories "déficience" et "incapacité". Les données sur les caractéristiques démographiques sont toujours essentielles puisqu'elles permettent à l'utilisateur de trouver quels sont les segments de la population qui font partie des différentes catégories. On peut comprendre facilement l'importance de la question de l'emploi puisque c'est la situation vis-à-vis de l'emploi qui détermine en grande partie si les personnes souffrant d'une incapacité peuvent vivre de façon autonome financièrement. L'enquête sur la population active permet déjà de recueillir énormément de données sur l'emploi. L'objet de l'enquête de suivi sera de rassembler des données relatives au choix restreint des emplois auxquels ont accès les personnes sélectionnées en raison de leur incapacité. En plus de procéder à l'analyse des données sur la population des personnes handicapées, il sera possible de comparer les caractéristiques de la main-d'œuvre handicapée avec celles de l'ensemble de la population active canadienne.



Le tableau 1 montre une ventilation par sexe du pourcentage de membres de l'échantillon sélectionnés à l'aide des tests 1, 2 et 3, et dans chaque section des tests 2 et 3. Ce qu'il convient de souligner dans ce tableau, c'est que le test 1 a permis de sélectionner seulement 5,6% de membres de l'échantillon, contre environ 16% de l'échantillon pour chacune des questions sur les AVQ. Comme la limitation fonctionnelle mesurée par l'accomplissement des activités de la vie quotidienne est un indicateur clé de l'invalidité, cela prouve que l'unique question posée dans le tests 1 n'est pas utile pour sélectionner toute la population cible.

Les pourcentages des personnes de l'échantillon sélectionnées à l'aide des tests 2 et 3 sont présentés dans la figure 1 où ils sont ventilés par âge. Dans les deux tests, la probabilité qu'une personne soit sélectionnée augmente suivant l'âge. Quoique à des degrés variables, la même relation vaut pour chacune des sections des tests 2 et 3, pour chacune des dix-sept AVQ et la majorité des affections chroniques (la sclérose en plaques, l'épilepsie, la paralysie cérébrale, la fibrose kystique et la dystrophie musculaire font exception à la règle).

Comme la limitation fonctionnelle constitue un bon indicateur d'invalidité, la question examinée dans l'analyse a été de déterminer comment inclure les AVQ dans le questionnaire de sélection afin que de savoir si on devait les inclure. L'enquête qui tenait compte du "recours à un appareil spécial" (test 2) a permis de dépister 15,5% des membres de l'échantillon, alors que l'enquête fondée sur le concept de "sans l'aide d'un appareil spécial" (test 3) a permis d'en sélectionner 16,0%. La différence entre ces chiffres n'est pas statistiquement significative. Pour les enquêtes et les interviews, la notion de "recours à un appareil spécial" a semblé naturelle et facile à comprendre; c'est pourquoi on a décidé de l'utiliser dans la version finale du questionnaire de sélection.

Comme la série complète des activités peut servir à mesurer le degré d'invalidité, les dix-sept questions sur les AVQ ont été incluses dans la version finale du questionnaire.

La section sur les appareils spéciaux du test 2 a permis de sélectionner 3,0% de l'effectif de l'échantillon. De ce pourcentage, 84,2% ont été repérés à l'aide de la section sur les AVQ. Ainsi, bien que ce genre de données soient intéressantes pour la base de données sur l'invalidité, la section sur les appareils spéciaux n'est pas un mécanisme de sélection efficace, en particulier si elle est jumelée avec la section sur les AVQ. Par conséquent, les questions sur les appareils spéciaux ne sont pas incluses dans la version finale du questionnaire de sélection.

Outre les activités de la vie quotidienne, la difficulté à accomplir une activité principale constitue aussi un important facteur d'invalidité. Le test 3 en a tenu compte par le biais des deux questions énoncées plus haut dans cette partie. Cependant, ces questions n'ont considéré que le point de vue de l'invalidité au travail. Parmi les personnes faisant partie de l'échantillon visé par le test 3, 13,1% ont été repérées à l'aide de ces deux questions (6,8% étaient des personnes ayant un emploi, 6,6% étaient sans emploi et 22,7% étaient hors de la population active). Bien qu'il n'y ait pas eu de problèmes manifestes provenant des données obtenues à l'aide de ces questions sur l'invalidité au travail, il y a eu certaines difficultés d'ordre opérationnel. En particulier, les questions ne convenaient pas toujours aux retraités. Par conséquent, la question relative à la difficulté d'accomplir une activité principale a été adaptée dans la version finale du questionnaire afin de mieux tenir compte des personnes hors de la population active.

La section sur les affections chroniques du test 3 a permis de dépister 23,9% des membres de l'échantillon visé par ce test. De ce nombre, 37,9% n'auraient pas été dépistés autrement. Cette section a posé deux problèmes principaux. Premièrement, la question s'est révélée difficile pour les enquêtes qui n'étaient pas certains de la nature de leur(s) affection(s). Deuxièmement, les données recherchées à l'aide du questionnaire de suivi sont généralement plus pertinentes pour les personnes qui souffrent effectivement d'une incapacité. Ainsi, les personnes dépendantes en raison d'une affection chronique mais qui n'ont actuellement aucune limitation fonctionnelle ni aucune difficulté à accomplir leur activité principale seraient interviewées dans le cadre du suivi et ne fourniraient probablement que peu de données utiles.

Compte tenu de ces problèmes, les affections chroniques ne sont pas utilisées comme critères de sélection dans le questionnaire final. Le handicap mental constitue une exception à cette règle. Cette affection seule est retenue comme élément de sélection, puisqu'il pourrait y avoir des personnes souffrant de handicap mental qui ne seraient pas dépistées à l'aide des AVQ, ni même à l'aide de la question sur la difficulté à accomplir l'activité principale.

Tableau 1  
Pourcentage de personnes sélectionnées dans chaque section de  
chaque questionnaire, par sexe

Section		Pourcentage de personnes sélectionnées	
		Hommes	Femmes
Test 1:	AVQ.....	5.5	5.7
Test 2:	Appareils spéciaux.....	14.7	16.2
Test 2:	AVQ.....	3.2	2.9
Test 2:	Activités limitées.....	5.9	6.1
Test 2:	.....	18.3	19.4
Test 3:	AVQ.....	15.4	16.7
Test 3:	Invalidité au travail.....	13.0	13.1
Test 3:	Affections chroniques.....	25.6	27.4
Test 3:	.....	29.8	31.2

La taille de l'échantillon du test 1 était d'environ 115,000 personnes. Pour chacun des tests 2 et 3, la taille de l'échantillon atteignait presque 38,000 personnes. Pour tous ces tests, les échantillons se composaient d'environ 49% d'hommes et 51% de femmes.

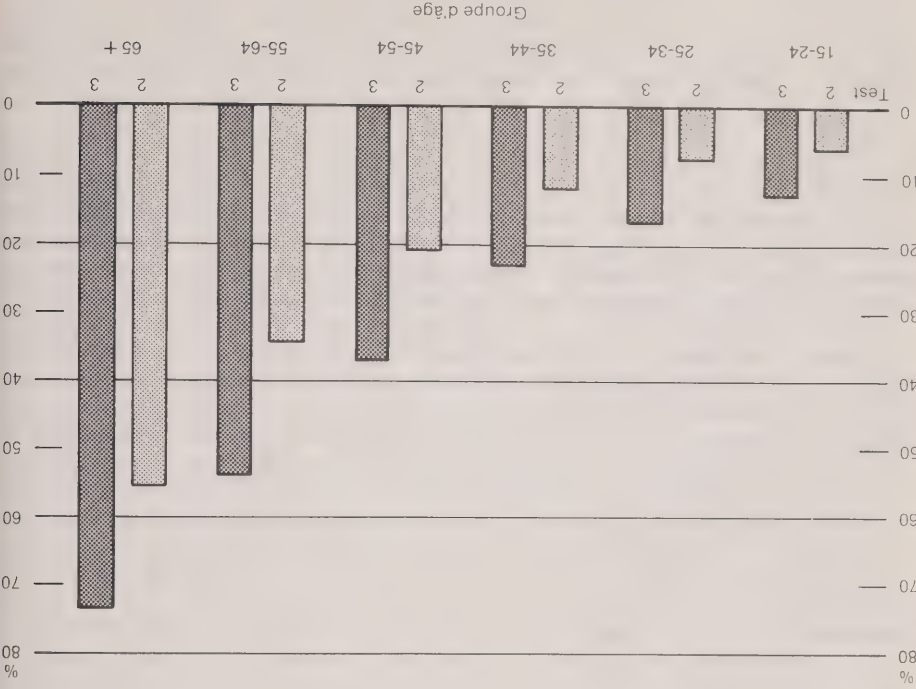


Figure 1. Pourcentage de personnes sélectionnées à l'aide des tests 2 et 3

## 6. MISE À L'ESSAI DES INSTRUMENTS DE SÉLECTION

Les trois tests de sélection de Statistique Canada comportaient divers aspects des démarches décrites plus haut. En outre, le questionnaire qui s'adressait aux personnes âgées de quinze ans et plus était différent de celui qui était conçu pour les moins de quinze ans. Aucune série approuvée d'AVQ n'a été établie pour les enfants. En fait, la plupart des enquêtes sur l'incapacité menées dans le passé ont exclu les enfants. Nous ne tiendrons compte ici que des personnes âgées de quinze ans et plus. Dans l'enquête sur la population active de novembre 1983, on a posé à chaque enquête la question suivante : "... souffre-t-il (elle) d'une invalidité ou d'un handicap d'une durée effective ou prévue de six mois ou plus?" C'est ce qu'on a appelé le test 1. Les personnes sélectionnées à l'aide du test 1 sont celles qui ont répondu "oui" à cette question.

Les autres méthodes de sélection ont été mises à l'essai à l'aide de deux questionnaires différents, et chacun a été utilisé dans le cadre de l'enquête sur la population active de janvier 1983. Ces questionnaires ont été appelés tests 2 et 3.

Le test 2 renfermait les sections suivantes : une liste d'appareils spéciaux, une liste d'AVQ et une question sur la difficulté à accomplir une activité : "Y a-t-il d'autres affections ou problèmes de santé qui empêchent ... d'accomplir ses activités quotidiennes normales à un emploi, à l'école ou à la maison ou qui le (la) limitent dans ces activités? Veuillez ne déclarer que les difficultés qui devraient durer plus de six mois?" Les personnes qui ont déclaré utiliser au moins un de ces appareils spéciaux ou éprouver de la difficulté à accomplir au moins une des AVQ ou qui ont répondu "oui" à la question sur les activités ont été sélectionnées.

Le test 3 comprenait une liste d'AVQ, une liste d'affections chroniques et les deux questions suivantes sur la difficulté à accomplir un travail : "... est-il (elle) limité(e) dans le genre ou la quantité de travail qu'il(elle) peut faire à son présent emploi ou entreprise en raison d'une affection physique ou d'un problème de santé?" (Cette question s'adressait uniquement à une personne ayant un emploi) et "Une affection physique ou un problème de santé empêche-t-il ou limite-t-il ... le genre ou la quantité de travail qu'il (elle) pourrait faire à n'importe quel emploi ou entreprise?" Les personnes qui ont déclaré éprouver de la difficulté à accomplir au moins une AVQ, ou souffrir d'au moins une des affections chroniques, ou qui ont répondu "oui" à l'une ou l'autre des deux questions sur le travail ont été sélectionnées.

Les listes des appareils spéciaux, des affections chroniques et des AVQ utilisées dans ces tests sont données en annexe. Il faut souligner que les AVQ utilisées sont les mêmes dans les tests 2 et 3, quoique légèrement modifiées par rapport aux AVQ de la liste de l'OCDE. En outre, chaque test applique de façon différente la liste des AVQ utilisées. Ainsi, le test 2 inclut l'utilisation d'appareils spéciaux pour accomplir les activités, alors qu'on n'en tient pas compte dans le test 3. Les méthodes de sélection proposées ne permettent pas d'évaluer si la population cible est définie de façon exacte, à moins qu'elles ne soient appliquées à une population témoin. Cela n'a pas été fait dans le cas de l'enquête canadienne.

Tous les membres des ménages choisis pour l'enquête sur la population active de novembre 1982 ont été soumis au test 1. En janvier 1983, on a administré les tests 2 et 3 à deux groupes de renouvellement. Ces groupes de renouvellement ont été choisis parmi les ménages qui faisaient partie de l'échantillon en novembre 1982, de façon à faciliter la comparaison des résultats du test 1 avec ceux des tests 2 et 3.

### 6.1 Principaux résultats des tests pilotes

Le but principal de l'analyse des tests était de déterminer la série de questions qui serait la plus utile pour sélectionner les personnes qui font partie de la population cible. Un autre facteur examiné avait trait à la méthode de sélection qui n'augmenterait pas indûment le fardeau de réponse ou le coût de l'enquête. En outre, les sources d'erreurs d'observation devenues évidentes durant l'analyse ont été prises en note, de sorte que, par exemple, les méthodes d'enquête ou la conception du questionnaire ont pu être corrigées. L'exposé suivant présente les principaux résultats de cette analyse.



La collecte des données pour l'enquête sur les personnes atteintes d'une ou de plusieurs incapacités s'est faite en deux étapes. Dans un premier temps, toutes les personnes de tous les ménages compris dans l'échantillon de l'EPA ont été soumises à un processus de sélection, exception faite du sixième de l'échantillon pour lequel il s'agit du premier mois de l'enquête. Un questionnaire de sélection a servi à identifier les personnes susceptibles de présenter de l'intérêt pour les fins de l'enquête. Le mode de collecte des données a été le même que celui qui est utilisé pour l'EPA. Toutefois, on a demandé aux enquêteurs d'obtenir autant que possible des interviews directes avec les personnes visées (et non pas par l'entremise d'un tiers) même s'ils ont été priés de répondre à une autre série de questions dans le cadre d'une enquête de suivi. Ces interviews se sont déroulées environ une semaine après l'interview de sélection. Il s'agissait exclusivement d'interviews directes, c.-à-d. qu'on a accepté uniquement les réponses des intéressés. Cette seconde série de questions devait permettre de recueillir les données jugées importantes à la suite des consultations avec les utilisateurs.

L'élaboration de l'enquête s'est déroulée de la façon suivante. Trois projets de questionnaire de sélection ont été mis à l'essai en novembre 1982 et en janvier 1983. (Des explications détaillées sur ces enquêtes sont données plus loin.) Le questionnaire de sélection définitif a été conçu d'après les résultats de ces essais. Deux enquêtes complètes comprenant un questionnaire de sélection et un questionnaire détaillé, tel qu'il est décrit ci-dessus, ont été menées en octobre 1983 et en juin 1984.

### 5. DÉMARCHES EMPRUNTÉES EN MATIÈRE DE SÉLECTION - AUTRES ENQUÊTES

Dans une première étape de conception d'un questionnaire de sélection, on a examiné les expériences d'autres groupes qui avaient déjà mené des enquêtes sur les personnes souffrant d'une incapacité.

Dans bon nombre des enquêtes, la question est abordée sous l'angle des "activités de la vie quotidienne" (AVQ). Il s'agit d'une série d'activités que toute personne est tenue d'accomplir dans la vie de tous les jours. Bien que "la meilleure" série d'activités à utiliser ne fasse pas l'unanimité, celle qui a été définie en 1978 par l'Organisation pour la coopération et le développement économiques (OCDE) (McWinnie (1980)) a été appliquée dans plusieurs pays; voir Klaukka (1981), Mizrahi et Mizrahi (1981), Raymond, Christie et Clemence (1981), Van Son-beek (1981), Wilson and McNeil (1981).

Comme la capacité d'une personne à accomplir une AVQ peut dépendre de l'utilisation d'un appareil spécial, un membre artificiel par exemple, une liste d'appareil spécial pourrait servir d'instrument de sélection.

Une autre façon de repérer les personnes concernées fait appel au concept de la difficulté à accomplir une activité principale. Si une personne est limitée dans son activité principale (c'est-à-dire travail, études, foyer), elle est sans doute atteinte d'une forme quelconque d'incapacité. Ce concept a été utilisé aux États-Unis dans le cadre d'un essai préliminaire pour une enquête sur les personnes souffrant d'une incapacité (1980) et pour la Health Interview Survey, qui est une enquête annuelle, de même qu'au Canada dans le cadre de l'Enquête Santé Canada (1978-1979).

Une liste d'affections chroniques pourrait être utile comme instrument de sélection, car il y a dans la population cible des personnes atteintes de maladies chroniques qui peuvent ne pas être repérées selon la méthode faisant appel au concept des AVQ ou selon celle fondée sur la notion de l'activité principale limitée, si elles sont atteintes de maladies qui entraînent une incapacité périodique.

Enfin, une autre façon d'enquêter sur l'incapacité serait de poser une seule question liée à la perception de soi, par exemple : "Avez-vous des incapacités ou des handicaps physiques ?".

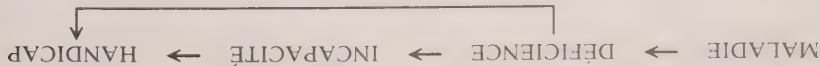
Lorsqu'on définit une population cible, on doit prévoir une mesure de la gravité, ne serait-ce qu'impartiellement. La question de la durée de l'invalidité doit être abordée explicitement. Si l'on tenait compte de toutes les incapacités, y compris celles qui découlent des maladies aiguës de durée relativement restreinte, on toucherait un fort pourcentage de la population, ce qui serait contraire à l'esprit de la recommandation contenue dans *Obstacles*. Toutefois, restreindre la population aux seules personnes atteintes d'incapacité permanente équivaut à esquiver le problème et à ne pas faire cas des besoins des personnes atteintes d'une incapacité de longue durée mais non chronique.

Comme nous l'expliquons dans la section suivante, les données ont été recueillies à l'aide de questions supplémentaires de l'enquête sur la population active (EPA). En plus des contraintes normales de l'EPA, le recours à cet instrument comportait une limite de taille: la population cible de l'enquête sur les personnes souffrant d'une incapacité ne devait pas comprendre les affections mentales. Par exemple, elle excluait les maladies telles que l'amnésie, les névroses et les phobies mais tenait compte des déficits de l'intellect tels que la déficience mentale et la dyslexie. On estimait, en effet, que le fait de demander ce genre de renseignement pouvait être très délicat et susciter des réactions négatives pouvant nuire aux objectifs principaux de l'enquête sur la population active.

Bref, la population cible choisie pour les tests de l'enquête sur les personnes souffrant d'une incapacité comprend toutes les personnes ayant une ou plusieurs incapacités physiques (non liées au comportement), ou des difficultés d'apprentissage (déroulant de troubles au niveau de l'intellect, de l'attention, des fonctions psychomotrices et du langage), d'une durée réelle ou prévue d'au moins six mois. Elle englobe également les personnes souffrant de maladies chroniques et dégénératives qui risquent fort d'engendrer des déficiences qui se traduisent par une incapacité physique. En outre, les contraintes normales de l'EPA font que les personnes placées en établissement s'en trouvent exclues.

#### 4. COLLECTE DES DONNÉES

Le problème est de traduire les définitions précédentes en une série de questions qui identifient les individus de la population cible dans l'ensemble de la population. Cela suppose qu'on se fixe pour objectif de recueillir des éléments d'information sur ceux qui ont une forte probabilité d'être atteints d'une ou de plusieurs incapacités selon l'acceptation la plus large du terme, tout en s'assurant que le nombre de personnes visées ne dépasse pas certaines limites raisonnables. La première option envisagée a été l'inclusion de questions dans le questionnaire du recensement de dix à trente nouvelles questions, ce qui était évidemment impossible. La deuxième option prise en compte a été l'inclusion dans ce questionnaire d'un nombre limité de questions qui se seraient adressées aux personnes atteintes d'une ou plusieurs incapacités. Pour satisfaire la demande de données, il aurait fallu une enquête de suivi. En fait, c'est la deuxième option qui a été retenue, et des questions ayant trait à l'incapacité seront incluses dans le questionnaire du recensement de la population de 1986. Cela permettra d'obtenir des estimations pour les petites régions. Les résultats de l'enquête de suivi ne pourront toutefois pas être connus avant 1988 ou 1989. Étant donné que la demande de données était forte et que cette demande visait principalement la production d'estimations nationales, l'option recensement et enquête de suivi a été jugée insuffisante. Pour répondre aux besoins courants de données, deux solutions ont été envisagées. La première possibilité consistait à mener, sur une base permanente ou périodique, une enquête auprès des ménages qui serait semblable à l'Enquête Santé Canada et qui aurait pour objet de recueillir des données sur les caractéristiques des personnes invalides et handicapées. Une telle enquête pourrait avoir recours aux méthodes connues expressément pour la collecte de données sur les personnes souffrant d'une incapacité. Des ressources limitées ont toutefois rendu cette solution irréalisable. La deuxième solution consistait à utiliser les questions supplémentaires de l'EPA. Pour des raisons de commodité et de coût, c'est cette méthode qui a été choisie. Accessoirement, cette méthode permet de relier directement, pour une même personne handicapée, les données ainsi obtenues sur son incapacité et les données ayant trait à sa situation de travail qui sont recueillies par l'EPA.



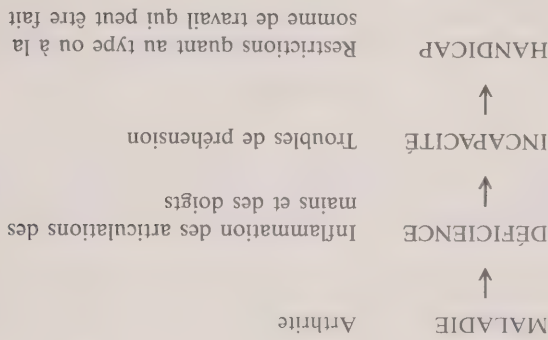
Ces termes se définissent ainsi (Organisation mondiale de la santé (1980)):

**Déficience:** Dans le contexte de l'expérience de la santé, la déficience correspond à toute perte ou altération d'une structure ou fonction psychologique, physiologique ou anatomique. Elle est caractérisée par une perte ou une anomalie transitoire ou permanente, et elle comprend l'existence ou l'avènement d'une anomalie, d'un défaut ou d'une perte d'un membre, d'un organe, d'un tissu ou d'une autre structure de l'organisme, dont les mécanismes liés au fonctionnement mental. Une déficience est la manifestation extérieure d'un état pathologique et, en règle générale, est le reflet d'un dérèglement organique.

**Incapacité:** Une incapacité correspond à toute réduction (résultant d'une déficience) partielle ou totale d'accomplir une activité d'une façon ou dans les limites considérées normales pour un être humain.

L'incapacité est caractérisée par des excès ou des lacunes par rapport à l'activité normale, et peut être temporaire ou permanente, réversible ou irréversible et progressive. Elle peut être la conséquence directe d'une déficience ou peut représenter pour l'individu une façon de réagir, surtout sur le plan psychologique, à une déficience physique, sensorielle ou autre. Il s'agit de l'extériorisation d'une déficience, et, à ce titre, traduit un dérèglement chez l'intéressé.

**Handicap:** Un handicap est un désavantage pour un individu donné résultant d'une déficience ou d'une incapacité, et qui limite ou interdit l'accomplissement d'un rôle normal (en rapport avec l'âge, le sexe, les facteurs sociaux et culturels). Le handicap est rattaché à l'importance accordée à la situation ou l'expérience vécue par un individu lorsque celles-ci s'écartent de la norme. Il se caractérise par un décalage entre la situation ou les réalisations d'un individu et ce que lui-même ou son entourage attendent de lui. Ainsi, un handicap apparaît comme la socialisation de la déficience ou de l'incapacité et montre les conséquences que peuvent avoir une déficience et une incapacité sur la vie culturelle, sociale, économique et le milieu d'un individu. Voici un exemple qui aidera à mieux expliquer ces définitions:



### 3. POPULATION CIBLE

Idealement, la population cible devrait comprendre toutes les personnes au Canada qui ont une incapacité au sens de la définition donnée plus haut et sous réserve des limites relatives à la gravité et à la durée de l'incapacité. La gravité peut être considérée du point de vue de la somme, c'est-à-dire dans quelle mesure la personne est-elle atteinte, ou encore du point de vue de l'incapacité proprement dite, à savoir dans quelle mesure l'incapacité en question est-elle grave.



# Méthode d'enquête sur les personnes souffrant d'une incapacité à l'aide de questions supplémentaires de l'enquête sur la population active<sup>1</sup>

D. DOLSON, P. GILES, et J.-P. MORIN<sup>2</sup>

## RÉSUMÉ

Pour répondre à la demande de données sur les personnes souffrant d'une incapacité au Canada, Statistique Canada a mis en oeuvre un programme visant à constituer une base de données sur l'incapacité physique. À cette fin, le Bureau a utilisé les données recueillies dans des suppléments de l'enquête sur la population active du Canada, à l'automne de 1983 et au printemps de 1984, et il inclura des questions sur ce sujet dans le questionnaire du recensement de la population de 1986. Ce document présente une analyse générale du cadre et du contenu de l'enquête. On compare également les divers processus de sélection utilisés dans les enquêtes menées par Statistique Canada en novembre 1982 et janvier 1983, de même que les résultats obtenus dans chaque cas.

MOTS CLÉS: Incapacité; questionnaire de sélection; activités de la vie quotidienne.

## 1. INTRODUCTION

Le 23 mai 1980, le gouvernement canadien instituait le Comité parlementaire spécial concernant les invalides et les handicapés. En février 1981 (l'Année internationale des personnes handicapées), le Comité publiait son rapport intitulé *Obstacles* [4]. Voici un extrait de la recommandation n° 113 :

"Que le gouvernement fédéral demande à Statistique Canada d'accorder une très haute priorité à l'élaboration et à la mise en oeuvre d'une stratégie à long terme visant à constituer une base de données sur les personnes souffrant d'une incapacité au Canada, et ce à l'aide d'enquêtes menées auprès de la population et de données tirées de divers programmes."

Désireux de donner suite aux recommandations contenues dans le rapport, le gouvernement a donc chargé Statistique Canada d'entreprendre une enquête auprès des personnes souffrant d'une incapacité. Le présent document porte sur les enquêtes menées auprès des personnes souffrant d'une incapacité dans des suppléments de l'enquête sur la population active du Canada (EPA), en octobre 1983 et en juin 1984, et sur les tests administrés en novembre 1982 et en janvier 1983.

## 2. DÉFINITIONS

Le Comité parlementaire spécial a appliqué les définitions établies par l'Organisation mondiale de la santé (OMS) McWinnie (1980). Ces définitions découlent d'un modèle qui met l'accent sur les conséquences de la maladie et tient compte des phénomènes suivants liés à la maladie.

<sup>1</sup> Cet article est une version combinée de deux articles intitulés: "Méthode d'enquête sur les personnes handicapées à l'aide de questions supplémentaires de l'enquête sur la population active du Canada" par P. Giles et D. Dolson et "Initiative canadienne en matière de sélection des personnes handicapées dans le cadre d'une enquête sur les ménages" par D. Dolson, P. Giles et J.-P. Morin. Ces articles ont été présentés à la réunion de l'ASA à Toronto en 1983.

<sup>2</sup> P. Giles, Division des méthodes d'enquête pour les entreprises, D. Dolson et J.-P. Morin, Division des méthodes d'enquête pour les institutions et l'agriculture, Statistique Canada, Parc Tunney, Ottawa, Ontario, Canada K1A 0T6.

chaque  $k$ ) unités du 1<sup>er</sup>, 2<sup>ème</sup>, ...,  $k$ <sup>ème</sup> domaine ont été sélectionnées. Pour chaque  $k$ , la taille  $\xi_k$  de l'échantillon du  $k$ <sup>ème</sup> domaine est une variable aléatoire et  $P_{\lambda}(\xi_k \geq m_k) = 1$  (où  $m = (M_1, \dots, M_k, \dots, M_p, E_p)$  sont respectivement les opérateurs de probabilité et d'espérance mathématique) parce que, dès que le nombre de tirages dans le  $k$ <sup>ème</sup> domaine a été complètement exécuté, il peut falloir poursuivre l'échantillonnage pour faire les tirages nécessaires dans les autres domaines, ce qui peut accroître le nombre d'unités tirées du  $k$ <sup>ème</sup> domaine. Le moyen  $\bar{x}_k$ , des unités de l'échantillon qui proviennent du  $k$ <sup>ème</sup> domaine est un ESB de  $\mu_k$  et a pour variance  $S^2[E_{\lambda}(\bar{x}_k) - 1/m_k]$ ,  $1 \leq k \leq t$ .

Des essais numériques (qu'on n'examine pas dans cet exposé parce qu'ils semblent déborder le cadre de cet article) révèlent que, pour un coût donné, l'estimateur de la méthode d'échantillonnage inverse est plus efficace que celui de l'BASSR avec un nombre fixe de tirages. Toute fois, dans l'échantillonnage inverse de plusieurs domaines, les expressions algébriques détaillées sont très complexes et rendent difficile une comparaison analytique comme celle résumée à la section précédente, pour cette raison, nous en faisons grâce au lecteur.

## REMERCIEMENTS

Les auteurs remercient l'arbitre pour ses observations constructives qui ont permis d'améliorer une version préliminaire de cet article.

## BIBLIOGRAPHIE

- Haldane, J.B.S. (1945). On a method of estimating frequencies. *Biometrika*, 33, 222-225.
- Rao, J.N.K. (1975). Analytical studies of sample survey data. *Survey Methodology*, 1, 1-76.
- Samford, M.R. (1962). Methods of cluster sampling with and without replacement for clusters of unequal sizes. *Biometrika*, 49, 27-40.

Le biais de  $\mu^*$  est  $-\mu P_M^{(c)} = 0$  (précisons que, si  $d \geq N - r + 1$ , on sait alors que  $M \in \mathcal{M}$  et que  $\mu^*$  se ramène à l'ESB défini plus haut) et on peut démontrer la propriété suivante de l'erreur quadratique moyenne de cet estimateur

$$EQM_{\mu^*} = S^2 \sum_{i=1}^n (1/\mu_i - 1/M) P_M^{(c)} = a + \mu^2 P_M^{(c)} = 0. \quad (3.6)$$

Il est difficile d'effectuer une comparaison analytique directe entre (3.5) et (3.6), mais des exemples numériques comme les deux présentés plus bas indiquent que, dans la plupart des cas pratiques, la valeur de (3.5) est moins élevée que celle de (3.6), ce qui signifie que notre estimateur est supérieur même si l'estimateur de l'EASSR avec un nombre fixe de tirages peut renfermer in biais.

**Exemple 3.1.** Les données suivantes sont les notes (en pourcentage) de tous les étudiants qui ont passé l'examen de baccalauréat en statistique de l'Indian Statistical Institute (ISI) au cours de cinq années scolaires consécutives jusqu'à 1984.

68	80	80	72	87	71	55	75	85	52	82
76	73	54	57	51	56	48	73	54	76	69
87	81	68	74	58	56	71	66	69	81	59
65	83	79	72	50	44	65	61	57	50	73
85	87	64	70	48	58	61	53	56	62	61
74	62	56	62	58	58	66	70	80	74	80

Supposons qu'on veuille estimer, à partir d'un échantillon, la note moyenne des étudiants qui se sont classés au premier rang (ceux qui ont obtenu une note de 60% ou plus). Dans ce problème,  $N = 66$ ,  $M = 44$ ,  $\mu = 44$ ,  $\mu = 73.1818$ , et  $S^2 = 61.6871$ . Dans l'EASSR avec un nombre fixe de tirages inverse est  $d(M + 1)/(N + 1) = 6.72$  et la variance (3.5) est égale à 8.8792 pour  $m = 6$  et à 7.4105 pour  $m = 7$ ; les gains d'efficacité par rapport à l'échantillonnage avec un nombre fixe de tirages sont de 8.08% et 29.50% respectivement.

**Exemple 3.2.** Voici un exemple un peu différent des problèmes habituels. Supposons qu'on veuille estimer la moyenne des nombres premiers compris dans les soixante premiers nombres naturels. Dans ce problème,  $N = 60$ ,  $M = 18$ ,  $\mu = 24.5$ , et  $S^2 = 350.1471$ . Dans l'EASSR avec un nombre fixe de tirages où  $d = 7$ , la valeur de (3.6) est 205.4654. Pour résoudre ce même problème par l'échantillonnage inverse, on doit faire  $m = 2.18$  tirages; pour  $m = 2$ , la variance (3.5) vaut 155.6209, ce qui correspond à un gain d'efficacité de 32.03%.

## 4. CONCLUSION

Le modèle décrit plus haut se limite au calcul d'une estimation pour un seul domaine. Dans les enquêtes menées sur une grande échelle, il est souvent nécessaire d'appliquer des estimateurs à plusieurs domaines, et le modèle peut être modifié pour combler cette lacune. Soit  $t$  le nombre de domaines, dont la taille,  $M_k$ , est inconnue et dont l'espace des paramètres est  $\mathcal{M}_k = \{r_k, R_k\}$ , où  $r_k$  et  $R_k$  sont connus ( $1 \leq k \leq t$ ). Soit  $\mu_k$  et  $S_k^2$  la moyenne et la variance de la population pour une variable particulière dans le  $k^{\text{ème}}$  domaine. L'échantillon est choisi selon un plan d'échantillonnage inverse fondé sur la loi hypergéométrique généralisée, c'est-à-dire qu'on continue l'EASSR inverse jusqu'à ce qu'au moins  $m_1, m_2, \dots, m_t$  ( $m_k \leq r_k$  pour

Pour  $0 \leq k \leq d$ , définissons  $t_k = \sum_{i \in A} (X_i^{(k)})^k$ ,  $\Sigma_k$  étant la somme obtenue pour tous les échantillons contenant exactement  $k$  unités du domaine  $A$ . On constate qu'aucune valeur de  $x_j$  n'intervient dans le calcul de  $t_0$ .  
 Si  $d = N - r$ ,  $\omega = \{N - d, N - d + 1, \dots, N\}$ . Supposons que  $M = N - d + j$  ( $0 \leq j \leq d$ ). Il y a alors  $\binom{N}{j}$  manières de choisir les unités appartenant au domaine  $A$ . Or, le système (3.1) comprend  $\binom{N}{j}$  équations. La somme des équations du système (3.1) pour toutes les manières de choisir les unités de  $A$  est

$$(3.2) \quad \sum_{j=0}^d a_j t_j^w = \frac{N - d + j}{\binom{N-1}{d-j} T},$$

où, pour  $0 \leq j \leq w \leq d$ ,  $a_j = \binom{N-w}{j} \binom{w}{d-j}$  si  $N - d \geq w - j$ ; et 0 autrement. L'équation (3.2) permet d'exprimer la solution des  $t_j$  sous la forme suivante:

$$(3.3) \quad t_j^w = \binom{N}{d} \binom{d}{w} \frac{T}{N^w} \quad 0 \leq w \leq d,$$

et la validité de (3.3) repose sur le fait que

$$\sum_{j=0}^w \binom{N}{d} \binom{d}{w} \binom{w}{j} \binom{w-j}{d-j} = \binom{N}{d} \binom{d}{w} \binom{w}{d-j}.$$

En particulier, la formule (3.3) donne comme résultat  $t_0 = N \binom{N}{d} T$ . Mais  $t_0$  n'est pas indépendant des  $X_j$  dans cette expression, ce qui contredit un raisonnement fait plus haut. La conclusion de la preuve du théorème 3.1 est donc nécessairement vraie, et ce théorème est vérifié. Essentiellement, pour estimer  $\mu$  sans biais dans une enquête par EASSR d'un nombre fixe d'unités ( $d$ ), il faut que  $d \geq N - r + 1$ , mais  $d$  peut être trop élevé (surtout si  $r$  est faible) pour que ce genre d'enquête soit praticable. Même si  $d \geq N - r + 1$ , on peut démontrer que l'EASSR n'est pas le contexte de l'EASSR est moins efficace que l'estimateur proposé à la section précédente quand le coût de traiter des observations est égal.

Par exemple, supposons que  $d \geq N - r + 1$  et définissons la variance suivante:

$$(3.4) \quad V_M(\hat{\mu}) = S^2 \left[ E_M(V) - \frac{1}{M} \right].$$

Dans l'échantillonnage inverse, le système d'équations (2.1) indique que le nombre théorique de tirages est  $m(N + 1)/(M + 1)$ . Pour que ce genre d'échantillon soit comparable à un autre échantillon fondé sur un nombre fixe ( $d$ ) de tirages, ces deux nombres de tirages doivent être égaux. Autrement dit, il faut que  $m = d(M + 1)/(N + 1)$ , expression qu'on peut introduire dans l'équation de variance du théorème 2.1 obtenir

$$(3.5) \quad V_M(\bar{y}) = S^2 \left[ \frac{d(M+1)}{N+1} - \frac{1}{M} \right].$$

Étant donné que

$$E_M(c^{-1}) > [E_M(c)]^{-1} = \frac{dM}{N} > \frac{d(M+1)}{N+1},$$

il s'ensuit que la variance (3.4) dépasse la variance (3.5), ce qui prouve l'efficacité supérieure de l'EASSR pour l'échantillonnage inverse.

Il est également intéressant de comparer ces deux techniques d'échantillonnage quand l'estimateur de  $\mu$  dans l'EASSR avec un nombre fixe ( $d$ ) de tirages peut renfermer un biais. Dans cette dernière méthode, l'estimation (par le quotient) habituelle de  $\mu$  est [voir, par exemple, Rao (1975)]

$$\mu^* = c^{-1} \sum_i X_i \quad \text{si } c > 0$$

$$= 0 \quad \text{si } c = 0$$



**Observation.** À l'aide du système (2.1) et du lemme 2.1, on peut calculer  $V_M^M(\mathcal{N})$  et un ESB de cette variance. Les expressions algébriques de ces grandeurs sont faciles à évaluer numériquement dans  $n$  importe quelle situation pratique, mais on les omet de cet exposé parce qu'elles sont un peu compliquées.

Dans les raisonnements présentés plus bas,  $S^2 = (M - 1)^{-1} \sum_{i=1}^M (x_i - \mu)^2$ ,  $q(u)$  et  $\ell(u)$  sont des ESB de  $M^{-1}$  et  $M^2$  respectivement (estimateurs définis par le système (2.2)),  $\Sigma^{-1}$  représente la sommation des valeurs correspondant à chacune des unités du sous-ensemble  $A$  incluses dans l'échantillon,  $\bar{x} = m^{-1} \Sigma^{-1} X_i^t$ ,  $Z = m^{-1} \Sigma^{-1} X_i^t$  et  $s^2 = (m - 1)^{-1} \Sigma^{-1} (X_i^t - \bar{x})^2$ .

**Théorème 2.1.** Un ESB de  $\mu$  est  $\bar{x}$ , qui a pour variance  $V_M^M(\bar{x}) = S^2(\frac{1}{m} - \frac{1}{M})$ . Un ESB de  $V_M^M(\bar{x})$  est  $v(\bar{x}) = s^2(m^{-1} - q(u))$ .

**Preuve.** Ce théorème est facile à démontrer et, pour cette raison, on en omet la preuve.

**Théorème 2.2.** (i) Un ESB de  $T$  est  $\bar{T}$  qui a pour variance

$$V_M^M(\bar{T}) = S^2(\frac{1}{M} - \frac{1}{M^2}) E_M^M(M^2) + \mu^2 V_M^M(M).$$

$$(ii) \quad v(\bar{T}) = T^2 - [\ell(u)(Z - s^2) + M^2] \text{ est un ESB de } V_M^M(\bar{T}).$$

**Preuve.** Le paragraphe (i) est facile à démontrer et, pour cette raison, on en omet la preuve. Pour (ii), il suffit de souligner le fait que

$$E[\ell(u)(Z - s^2) + M^2] = \mu$$

$$= \ell(u)(M^{-1} \sum_{i=1}^M X_i^t - s^2) + M^2 = \ell(u)(\mu^2 - M^{-1} S^2) + M^2.$$

Ainsi, on peut écrire:

$$E_M^M(v(\bar{T})) = E_M^M(T^2) - [M^2(\mu^2 - M^{-1} S^2) + M^2] = E_M^M(T^2) - T^2 = V_M^M(\bar{T}).$$

### 3. COMPARAISON ENTRE L'ÉASSR INVERSE ET L'ÉASSR AVEC UN NOMBRE FIXE DE TIRAGES

Cette section montre que, si l'objectif primordial est de réduire le biais autant que possible dans l'estimation de  $\mu$ , la méthode décrite plus haut est supérieure à l'ÉASSR avec un nombre fixe de tirages. Cette section montre ensuite que cette supériorité ne se démontre pas même si les estimateurs renferment un certain degré de biais.

Soit  $d$  un nombre fixe de tirages dans une enquête fondée sur l'ÉASSR ( $d$  étant déterminé d'une manière bien définie),  $\delta$  un échantillon prélevé selon cette méthode,  $\delta \cap A$  l'ensemble d'unités du domaine incluses dans  $\delta$  et  $C$  le nombre cardinal de  $\delta \cap A$ . Les symboles  $P_M^M$ ,  $V_M^M$  représentent les mêmes variables que celles qui sont définies aux sections précédentes. On obtient les résultats suivants:

**Théorème 3.1.** Il existe un ESB de  $\mu$  si et seulement si  $d \geq N - r + 1$ .

**Preuve.** Supposons que  $d \geq N - r + 1$ . Il s'ensuit que  $P_M^M[C = 0] = 0$ ,  $\forall M \in \mathcal{M}$  et que

$$t = C^{-1} \Sigma^{-1} X_i^t \text{ est un ESB de } \mu.$$

Pour démontrer que cette conclusion est nécessairement vraie, il suffit de prouver que, si  $d = N - r$ , il n'existe pas d'ESB de  $\mu$ . Définissons d'abord quelques notations. Soit  $j_1, \dots, j_d$  des unités différentes rangées en ordre ascendant qui sont tirées des  $1, \dots, N$ , éléments de  $\delta$  et dont  $k$  unités ( $0 \leq k \leq d$ ),  $i_1, \dots, i_k$  (rangées en ordre ascendant) appartiennent au sous-ensemble  $A$ . On peut donc définir  $\delta = (i_1, \dots, j_d)$ ,  $\delta' = (i_1, \dots, i_k)$  est  $\delta \cap A$  (ce qui signifie que  $k = 0 \Rightarrow \delta' = \emptyset$  et  $k = d \Rightarrow \delta' = \delta$ ) et  $X(\delta') = (X_{i_1}, \dots, X_{i_k})$ , la série de valeurs de  $X_i$  pour les unités de  $\delta'$ .

Si  $t$  est un ESB de  $\mu$ , on peut écrire  $t = \ell(X(\delta'))/\delta$  et il s'ensuit que

$$E_M^M(t) = \mu, \quad \forall X_1, \dots, X_M \in \mathcal{M}.$$

4. En pratique, on a souvent une idée générale de la valeur de  $M$ . Définissons l'espace des paramètres de  $M$  comme suit:  $\theta = \{r+1, \dots, R\}$ , où  $r \geq 1$  et  $R(\leq N)$  sont connus. Dans presque tous les cas réels,  $r$  est beaucoup plus grand que 1 et  $R$  est beaucoup plus petit que  $N$ . Soit  $X_1, \dots, X_N$  les valeurs de  $y_i$  des  $M$  unités de  $A$ . On veut des estimateurs de  $\mu = (\sum_{i=1}^N X_i)/M$ , et peut-être aussi de  $M$  et  $T = \sum_{i=1}^N X_i$ , pour un échantillon comprenant un nombre préétabli  $m (\leq r)$ , d'unités de  $A$ . On peut utiliser les formules de la variance de ces estimateurs, qui est exprimée plus bas en fonction de  $m$ , pour calculer une valeur acceptable de  $m$ . Il est commode d'écrire  $X_{N+1} = \dots = X_N = 0$  pour les unités de  $I$ , exclues du domaine  $A$ . Supposons qu'on choisit des unités par une série de tirages fondée sur l'EASSR jusqu'à ce qu'on dispose d'exactly  $m$  unités de  $A$ . Le nombre de tirages,  $u$ , est alors une variable aléatoire dont la distribution de probabilité,  $P_y(\cdot)$  (qui dépend du paramètre inconnu  $M$ ), est

$$P_y(u=n) = \binom{M}{D} \binom{m-1}{n-m} \binom{N-1}{n-1} \cdot \frac{N-m+1}{N-n+1} = g_{Mn} \quad (2.1)$$

(pour résumer)  $(m \leq n \leq D+m)$ ,

où  $D = N - M$ . Pour éviter de résoudre des problèmes triviaux, nous supposons dorénavant que  $m \geq 2$ , hypothèse raisonnable. Les résultats suivants découlent donc du plan d'échantillonnage inverse décrit plus haut.

**Lemme 2.1.** Toutes les fonctions paramétriques  $f(M)$  sont estimables sans biais.

**Preuve.** Soit  $h(u)$ , si on peut calculer cette fonction, un estimateur sans biais de  $f(M)$ . On peut donc écrire

$$f(M) = \sum_{n=m}^{D+m} h(n) g_{Mn}, \quad r \leq M \leq R. \quad (2.2)$$

Si on récrit en notation matricielle ce système de  $R - r + 1$  équations à  $N - r + 1$  inconnues,  $h(m), \dots, h(N - r + m)$ , le fait que  $g_{Mn} > 0$  ( $m \leq n \leq D + m, r \leq M \leq R$ ) signifie que le rang de la matrice de coefficients correspondante est égal au nombre de lignes qu'elle contient. Par conséquent, on sait pertinemment qu'il existe une solution, et le lemme 2.1 est donc prouvé. **Observation.** Si  $R = N$ , le nombre d'équations dans le système (2.1) est égal au nombre d'inconnues. Dès lors, la matrice de coefficients est non singulière et on peut calculer une estimation unique sans biais pour chacune des fonctions paramétriques  $f(M)$ . **Corollaire 2.1.** Un ESB de  $M$  qui dépend de la valeur de  $u$  est  $\hat{M}(u) = N(m - 1)/(u - 1) = \hat{M}$  (pour résumer cette équation). **Preuve.** Premièrement, comme on suppose que  $m \geq 2, u > 1$  avec certitude (quelle que soit la valeur de  $M$ ) et la fonction  $\hat{M}(u)$  est donc bien définie. Or:

$$E \left( \frac{u-1}{1} \right) = \sum_{n=m}^{D+m} \frac{n-1}{1} g_{Mn}$$

$$= \frac{M! D!}{(M-m)!(m-1)! N!} \sum_{n=m}^{D+m} \frac{(n-m)!(D-n+1)!}{(n-2)!(N-n)!} = \frac{M! D!}{(M-m)!(m-2)!(N-1)!} \cdot \frac{(M-m)!(m-1)! D!}{(M-m)!(m-2)!(N-1)!} = \frac{M}{N(m-1)}, \quad \forall M \in \mathcal{M}.$$

ce qui prouve le corollaire 2.1.



# Estimation sans biais des paramètres d'un domaine dans l'échantillonnage sans remise

## ARJIT CHAUDHURI ET RAHUL MUKERJEE<sup>1</sup>

### RÉSUMÉ

Soit une population finie de taille  $N$  comprenant  $M$  (inconnu) unités qui appartiennent à la catégorie  $A$  et constituent un domaine dont la moyenne est  $\mu$ . Cette étude décrit une méthode d'échantillonnage aléatoire simple sans remise (EASSR) dans laquelle on choisit un nombre préalable d'éléments du domaine. On propose également un estimateur sans biais pour  $\mu$ . Cet estimateur est supérieur à son homologue dans l'EASSR avec un nombre fixe de tirages, même si ce dernier renferme un biais. Le modèle proposé produit également des estimateurs sans biais de  $M$  et du total  $T$  pour l'ensemble du domaine visé.

MOTS CLÉS: Domaine d'estimation; échantillonnage aléatoire simple sans remise.

### 1. INTRODUCTION

Pour assurer une utilisation rationnelle des ressources et l'efficacité des estimateurs dans les enquêtes à grande échelle, il est souvent nécessaire de prélever un échantillon qui représente bien une certaine catégorie (appelons-la " $A$ ") d'unités possédant des caractéristiques recherchées. Par exemple, les clients qui commandent une enquête et les utilisateurs des données peuvent préciser qu'ils veulent des estimations calculées à partir d'un échantillon comprenant un nombre déterminé d'agriculteurs (i) qui emploient un engrais particulier, (ii) qui pratiquent une certaine méthode d'irrigation et de culture et (iii) qui sont prêts à répondre honnêtement à une série de questions; (2) de fabricants qui utilisent le fer et l'acier pour un même besoin; (3) de membres d'un ménage qui ont un niveau d'instruction donné, etc. En dépit du soin qu'on peut mettre à élaborer un plan d'échantillonnage approprié, il est souvent possible que la base de sondage ne soit pas exacte. Supposons qu'on dispose d'une base de sondage imparfaite contenant  $N$  unités, où  $N$  est bien au-dessus de  $M$ , le nombre réel d'unités dans la catégorie  $A$ . On doit donc chercher une technique d'échantillonnage permettant d'estimer la moyenne (ainsi que la valeur globale et la taille) du domaine composé des membres de la catégorie  $A$ . On aborde ce problème plus bas à l'aide d'un plan d'EASSR "inverse". Différents auteurs ont déjà consulté des plans d'échantillonnage inverses avec remise (voir Haldane (1945) et Sampford (1962), entre autres) pour estimer la proportion  $f = M/N$  d'éléments dans le domaine étudié. Rao (1975) a aussi présenté des estimateurs de la moyenne  $\mu$  du domaine, mais ce sont des estimations par le quotient et elles ne sont pas sans biais. La technique d'EASSR inverse décrite plus bas permet d'obtenir un estimateur sans biais de  $\mu$  qui est plus efficace que son homologue dans l'EASSR avec un nombre fixe de tirages, même si ce dernier renferme un biais.

### 2. MÉTHODE D'ÉCHANTILLONNAGE ET D'ESTIMATION

Soit une population  $I_n = (1, \dots, j, \dots, N)$  composée de  $N$  unités numérotées  $1, \dots, j, \dots, N$  qui correspondent aux valeurs  $y_1, \dots, y_j, \dots, y_N$ . Un nombre  $M$  (inconnu) de ces unités possèdent certaines caractéristiques spéciales et constituent un ensemble ou un domaine que nous appellerons

<sup>1</sup> Arjit Chaudhuri et Rahul Mukerjee, Indian Statistical Institute, 203, rue Barrackpore Trunk, Calcutta 700 035, India.



- HANSEN, M.H., HURWITZ, W.N., MARKS, E.S. et MAULDIN, W.P. (1951). Response Errors in Surveys. *Journal of the American Statistical Association*, 46, 147-190.
- HANSEN, M.H., MURWITZ, W.N. et BERSHAD, M. (1961). Measurement Errors in Censuses and Surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.
- HARTLEY, H.O. (1981). Estimation and Design for Non-sampling Errors of Surveys. Dans *Current Topics in Survey Sampling*, D. Krewski, R. Platek et J.N.K. Rao éditeurs. New York: Academic Press.
- HORVITZ, D.C. (1981). Response Error Research Issues in Health Surveys. *1981 Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 397-399.
- KIBLER, W.E. (1978). Le contrôle des erreurs d'enquête non dues à l'échantillonnage. Document non publié, Comité fédéral-provincial de la statistique agricole, Statistique Canada.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.
- KISH, L. et LANSING, J.B. (1954). Response Errors in Estimating the Value of Homes. *Journal of the American Statistical Association*, 49, 520-538.
- KROTKI, K. (1980). *Erreurs de réponse au recensement de la population et du logement de 1976*. Document de travail, Ottawa, Canada: ministère des Approvisionnement et Services.
- MARQUIS, K.H., MARQUIS, M.S. et POLICH, J.M. (1981). Survey Responses to Sensitive Topics. *1981 Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 339-341.
- NISSLETON, H. et BAILLAR, B.A. (1976). Measurement, Analysis and Reporting of Non-sampling Errors in Surveys. *Proceedings of the 9th International Biometric Conference*, 2, 201-322.
- PHILLIPS, J. (1978). 1979 Farm Expenditure Survey design and Estimation Procedures. Document de travail, Statistique Canada, Division des méthodes d'enquêtes-institutions et agriculture.
- STATISTIQUE CANADA (1979). *Recensement du Canada de 1976 - Agriculture - Évaluation de la qualité des données*, (n° 96-872 au catalogue). Ottawa, Canada: ministère des Approvisionnement et Services.
- STATISTIQUE CANADA (1980). *Recensement du Canada de 1976 - Qualité des données - série I: Sources d'erreurs: Couverture*, (n° 99-840 au catalogue). Ottawa, Canada: ministère des Approvisionnement et Services.
- STATISTIQUE CANADA (1982). *Recensement du Canada de 1981 - Agriculture*, (n° 96-901 au catalogue). Ottawa, Canada: ministère des Approvisionnement et Services.
- STATISTIQUE CANADA (1984). *Recensement du Canada de 1981 - Agriculture - Évaluation de la qualité des données*, (n° 96-918 au catalogue). Ottawa, Canada: ministère des Approvisionnement et Services.
- SUKHATME, P.V. et SEETH, G.R. (1952). Non-sampling Errors in Surveys. *Journal of the Indian Society of Agricultural Statistics*, 4, 5-41.
- TREMBLAY, V., SINGH, M.P. et CLAVEL, L. (1976). Methodology of the Labour Survey Re-interview Program. *Techniques d'enquête*, 2, 43-62.
- U.S. BUREAU OF THE CENSUS (1964). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Accuracy of Data on Population Characteristics as Measured by the CPS - Census Match*. Series ER60, n° 5, Washington, D.C.
- U.S. BUREAU OF THE CENSUS (1970). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Record Check Study of the Accuracy of Income Reporting*. Series ER60, n° 8, Washington, D.C.; U.S. Government Printing Office.
- U.S. BUREAU OF THE CENSUS (1982). *1978 Census of Agriculture. Volume 5: Special Reports - Part 3. Coverage Evaluation*, (AC78-SR-3), Washington, D.C.; U.S. Government Printing Office.

de l'évaluation sont en train d'être pris en considération dans la planification du recensement de 1986. Les erreurs de couverture et de réponse découlent de l'évaluation des données de 1976 montrant les aspects des méthodes de collecte et de traitement qu'il faut songer à modifier ment au recensement de l'agriculture. La comparaison des réponses fournies au recensement et aux enquêtes a également fait ressortir des problèmes dans les enquêtes. Il est possible qu'il y ait des améliorations soient apportées à l'enquête nationale sur les fermes, l'enquête probabilité annuelle qui a remplacé l'EDA et l'EDDA, en fonction des résultats de l'évaluation des données du recensement.

Un autre avantage de cette étude concerne les connaissances acquises sur l'appariement d'enregistrements pour les besoins d'une évaluation. L'expérience dans la comparaison de données d'enregistrements particuliers par ordinateur et manuellement pourrait être très utile pour d'autres projets d'appariement. La compréhension des problèmes qui se présentent dans ce genre d'étude, leur causes, leurs caractéristiques et leurs solutions pourraient permettre d'améliorer la méthodologie des évaluations ultérieures.

En résumé, le projet d'évaluation de la qualité des données du recensement de l'agriculture de 1981 confirme de nouveau l'utilité de l'appariement d'enregistrements dans l'étude de l'erreur non due à l'échantillonnage. Les analyses des comparaisons globales et détaillées ont produit des mesures de la qualité des données du recensement de 1981 et indiquent les variables dont les erreurs ont un effet prononcé sur les résultats. Le mandat de ce programme a été étendu de manière à inclure la préparation de mises en garde sur les erreurs qui peuvent exiger la modification des méthodes de dénombrement pour la planification du recensement de 1986 et des recensements subséquents de cette évaluation à également de ce problème qui peuvent fausser les résultats des enquêtes de comparaison, ce qui peut faciliter la planification des enquêtes agricoles futures. Enfin, cette étude a permis d'acquiescer une expérience et des connaissances précieuses dans l'utilisation de techniques d'appariement d'enregistrements pour l'évaluation de la qualité des données.

## REMERCIEMENTS

L'auteur aimerait remercier D. Royce, M.R. Dibbs, B.N. Chinappa, K. Thatcher, l'arbitre et les autres membres de Statistique Canada dont l'appui et les observations pertinentes ont été fort appréciées au cours de la rédaction de cet exposé.

## BIBLIOGRAPHIE

ANDERSEN, R., KASPER, J., FRANKEL, M.R. et coll. (1979). *Total Survey Error*. San Francisco: Jossey-Bass Publishers.

COCHRAN, W.G. (1977). *Sampling Techniques*, troisième édition. New York: John Wiley & Sons.

DEMING, W.E. (1944). On Errors in Surveys. *American Sociological Review*, 9, 356-369.

FAULKENBERY, D., et TORTORA, R.D. (1981). Non-sampling Errors in an Agriculture Survey. 1981 *Proceedings of the Section on Survey Research, Methods of the American Statistical Association*, 493-495.

FELLEGI, I.P. (1964). Response Variance and its Estimation. *Journal of the American Statistical Association*, 59, 1016-1041.

FELLEGI, I.P. (1973). The Evaluation of the Accuracy of Survey Results: Some Canadian Experiences. *International Statistical Review*, 41, 1-14.

GOSSLIN, J.-F., CHINNAPPA, B.N., GHANGURDE, P.D. et TOURIGNY, J. (1978). *Répartition de méthodes d'évaluation des erreurs dans les recensements et les enquêtes*, (n° 13-564 au catalogue). Ottawa, Canada: Statistique Canada.



Tableau 7

Données sur une exploitation particulière fournies au recensement et à une enquête

Nombre total de bovins et de vœux	Superficie		L'exploitation (acres)		En culture		Pâturage amélioré		Jachère		Autre Terre		Autre terre à bois		Améliorée	
	totale de	(acres)	En	culture	Pâturage	Jachère	Améliorée	terre à bois	Autre	Terre	Améliorée	non améliorée	terre	bois	Améliorée	non améliorée
Recensement	2,640	1,035	0	1,240	15	0	350	0	15,000	815						
Enquête	17,000	1,010	0	970	20	0	15,000	815								

rubrique "autre terre non améliorée". Les hypothèses découlant de ces cas particuliers ont en- suite été vérifiées pour déterminer si elles sont applicables à l'ensemble du fichier.

## 6. CONCLUSION

L'appariement du fichier du recensement de 1981 avec des fichiers d'enquêtes est un moyen puissant d'évaluer les erreurs de couverture et de réponse dans le recensement. On savait que les données d'enquêtes utilisées comme approximations des valeurs exactes renfermaient des erreurs et que des lacunes dans le mécanisme d'appariement avaient favorisé indûment ou empêché l'appariement de certains enregistrements. Malgré ces difficultés, le fichier d'enregistrements appariés s'est avéré une source fort utile pour l'étude de la qualité des don- nées. Des analyses de couples d'enregistrements appariés permettent d'approfondir les premiers résultats des comparaisons globales des estimations du recensement avec celles d'enquêtes in- dépendantes. Ces analyses facilitent également l'étude de problèmes précis parce qu'elles ren- dent possibles des examens détaillés d'aspects particuliers.

L'évaluation a révélé des faits importants sur le sous-dénombrement dans le recensement. Des analyses des enregistrements d'enquête qui n'ont pas pu être appariés ont démontré que les exploitations omises dans le recensement ont, en général, une superficie totale, du bétail et des ventes de produits agricoles au-dessous de la moyenne. Les résultats offrent donc des arguments concrets en faveur de l'hypothèse très répandue selon laquelle des exploitations qui ont une valeur économique et agricole marginale sont parfois omises dans le recensement. Des écarts étaient concentrés dans certaines catégories, qu'il existe un peu de confusion entre différentes catégories et que les écarts entre les réponses varient d'une région à une autre du Canada. Certaines modifications de la structure et du libellé des questionnaires ou des méthodes de collecte des données ont été proposées à la suite des résultats de cette évaluation. Quelques- unes des variations observées sont attribuables au fait qu'il est difficile de définir certaines catégories d'utilisation des terres parce que les notions de base ne sont pas assez claires. Des problèmes comme celui-ci, qui viennent de la confusion que les répondants éprouvent quant à la manière de classer les terres selon différentes catégories d'utilisation, ne seront peut-être jamais résolus définitivement. Il est toutefois très utile pour la planification des recensements futurs d'admettre l'existence de ce genre de problème et d'en étudier les caractéristiques et les effets sur les données.

L'évaluation des données de 1981 a eu des répercussions profondes sur le recensement. Con- formément à son objectif, elle a mis en évidence des lacunes dans la qualité des données du recensement de 1981. Un document (Statistique Canada, 1984) a été publié pour fournir aux utilisateurs des précisions sur la qualité des données et les prévenir des problèmes particuliers qui ont eu un effet marqué sur les données. Dans une perspective à long terme, les résultats

Tableau 6

Comparaison des réponses du recensement et de l'EDA-EDEA dans les enregistrements appariés - Nombre d'exploitations selon l'écart entre les réponses concernant la superficie des terres à bois et la superficie déclarée dans le recensement (nombre d'exploitations ayant des terres à bois<sup>b</sup>), Canada<sup>a</sup>, 1981

Écart entre la superficie de terres à bois dans le recensement (acres)	Pourcentage des fermes ayant des terres à bois	Superficie des terres à bois déclarée dans le recensement (acres)				
		Total	400	240	70	à ou plus

Total	3,177	149	709	3,106	1,598	285	247	9,271	100.9
plus de 500	—	—	—	—	—	—	36	36	0.4
251 à 500	—	—	—	—	—	33	42	75	0.8
151 à 250	—	—	—	—	65	25	17	107	1.2
51 à 150	—	—	—	50	294	45	32	421	4.5
1 à 50	—	55	229	1,288	492	53	30	2,147	23.2
0	—	26	165	481	151	21	21	865	9.3
— 50 à — 1	1,705	59	277	1,053	370	57	28	3,549	38.3
— 150 à — 51	715	5	30	167	145	24	19	1,105	11.9
— 250 à — 151	268	4	1	27	44	9	4	357	3.9
— 500 à — 251	248	0	4	24	21	12	7	316	3.4
moins de — 500	241	0	3	16	16	6	11	293	3.2

<sup>a</sup> Sans Terre-Neuve, le Yukon et les Territoires du Nord-Ouest.

<sup>b</sup> Inclut toutes les exploitations qui ont déclaré des terres à bois dans le recensement ou les enquêtes; exclut les exploitations dont l'enregistrement dans le recensement ou les enquêtes contient uniquement des données imputées à cause de la non-réponse.

À cause de ces différences, on a estimé que des superficies de terres à bois d'une valeur commerciale incertaine avaient peut-être été déclarées dans les questionnaires d'enquête mais ex-cusés du recensement. Il est également possible que des terres utilisées à certaines fins, selon les réponses données aux enquêtes, aient été classées différemment, par exemple dans la catégorie "autre terre non améliorée", au recensement. Ces hypothèses et d'autres continuent d'être étudiées à l'aide du fichier des enregistrements appariés pour rechercher les causes possibles des écarts observés entre les réponses.

Quand des totalisations sommatrices calculées à partir du fichier des enregistrements appariés n'ont pas permis de trouver les causes possibles des écarts observés, les analyses d'enregistrements ont permis de détecter des erreurs de classement dans différentes catégories. Une comparaison des réponses à des questions connexes dans le recensement et les enquêtes a permis de constater des erreurs de classement dans différentes catégories souvent fructueuses. Les enquêtes ont permis de constater des erreurs de classement dans différentes catégories.

Le tableau 7 montre quelques-unes des réponses d'une exploitation dont les enregistrements contiennent un des écarts les plus grands entre la superficie totale déclarée au recensement et dans une des enquêtes. Dans ce cas particulier, l'écart entre les deux superficies totales était surtout attribuable à l'interprétation de la catégorie "autre terre non améliorée"; les différences entre les réponses dans les catégories "en culture", "jachère" et "autre terre améliorée" n'étaient pas importantes. Il semble que la plus grande partie de la superficie déclarée sous la catégorie "autre terre non améliorée" dans une des enquêtes a été exclue de la réponse fournie lors du recensement. Des analyses du même genre ont été faites avec d'autres enregistrements pour trouver les exploitations qui pouvaient avoir déclaré différentes superficies sous la





enquêtes aient été jumelées à la mauvaise forme de recensement à cause de similarités dans le tant dans le calcul des totaux provinciaux relatifs aux produits ou aux superficies n'est pas ap- partie à la bonne forme de recensement, la comparaison des résultats risque d'être nettement faussée. Une étude détaillée d'un échantillon des enregistrements appariés a été entreprise pour vérifier la qualité de l'appariement par ordinateur, mais elle n'était pas terminée au moment de l'évaluation. Par conséquent, il a fallu tenir compte de l'effet possible d'appariements er- ronnés sur les résultats calculés à partir du fichier des enregistrements appariés.

La deuxième difficulté dans l'analyse des enregistrements a compliqué la comparaison des totaux obtenus dans le recensement et les enquêtes. Les enregistrements appariés étaient un sous- ensemble de l'échantillon non auto-pondéré de l'EDA et de l'EDBA parce qu'ils correspon- daient seulement aux exploitations qui ont pu être repérées dans le fichier du recensement. Il aurait été préférable de multiplier les données des enregistrements appariés par les facteurs d'ex- pansion appropriés pour produire des estimations pondérées. Cependant, les facteurs d'expan- sion de la base aréolaire ont été calculés à partir des réponses obtenues dans les enquêtes au sujet de l'utilisation des terres, et on ignorait si ces facteurs étaient valables pour les données du recensement dans le fichier des enregistrements appariés. Il était impossible de calculer des facteurs d'expansion pour les données du recensement parce que des données sur une des com- posantes de ces facteurs, la superficie consacrée à l'exploitation agricole dans chaque segment choisi, ont été recueillies uniquement dans les enquêtes. Comme on n'était pas certain qu'on pouvait appliquer les facteurs d'expansion des données aux données des enregistrements ap- parités du recensement, on a décidé de limiter l'analyse aux totaux non pondérés du recense- ment et des enquêtes calculés à partir du fichier des enregistrements appariés. (Une étude d'estima- tions pondérées calculées à partir du fichier des enregistrements appariés était en cours pendant la rédaction du présent exposé, mais aucune conclusion n'en est encore ressortie.)

Malgré les désavantages d'un échantillon non auto-pondéré et la possibilité d'appariements erronés, le fichier des enregistrements appariés s'est avéré un outil d'évaluation très utile. Quand une comparaison des totaux calculés pour les enregistrements appariés indiquait des écarts entre les résultats du recensement et ceux des enquêtes, des analyses détaillées étaient faites pour rechercher les causes des différences observées.

Pour illustrer l'emploi de totaux non pondérés dans l'analyse des enregistrements appariés, le tableau 5 présente des chiffres globaux concernant l'ensemble du Canada sur la superficie totale des exploitations agricoles et les catégories d'utilisation des terres. Ces résultats indiquent que les superficies déclarées dans le recensement sont moins grandes que dans les enquêtes pour toutes les catégories d'utilisation des terres sauf les rubriques "pâturage amélioré" et "autre terre améliorée". Les écarts relatifs les plus faibles entre les totaux du recensement et des en- quêtes sont observés dans des catégories comme celles des terres en culture, qui ont une grande valeur économique et sont clairement définies et rarement mal interprétées par les exploitants agricoles. En général, plus la valeur agricole ou économique d'un aspect de l'agriculture est faible, plus les écarts entre les chiffres du recensement et des enquêtes sont élevés. Les écarts relatifs les plus importants ont été constatés dans la catégorie "terre à bois". Pour montrer quelques-unes des techniques d'évaluation détaillée que l'appariement des enregistrements rend possibles, les résultats d'une analyse des réponses concernant les terres à bois sont décrits plus bas.

## 5.5 Comparaisons détaillées des enregistrements appariés

La comparaison des totaux des enregistrements appariés est un moyen de mesurer les erreurs globales dans les réponses, mais elle n'offre aucune précision sur la répartition et la cause des écarts observés. Par exemple, les différences entre les superficies de terre à bois sont-elles at- tribuables seulement à une poignée d'exploitations ou sont-elles réparties également entre tous les enregistrements? Est-ce que l'ampleur et le sens des différences entre les réponses varient en fonction des types d'exploitations ou des régions du pays? En comparant les données qui

Tableau 4

Répartition en pourcentage des estimations de l'EDA-DEDA relatives au nombre total de fermes et au nombre de fermes non apparées selon la superficie totale de la ferme et la valeur totale des produits agricoles vendus en 1980, 1981, Canada<sup>a</sup>

Variable	Estimation de l'EDA-DEDA du nombre total de fermes <sup>b</sup>		Estimation de l'EDA-DEDA du nombre de fermes non apparées <sup>c</sup>	
	Pourcentage	Pourcentage cumulatif	Pourcentage	Pourcentage cumulatif
Superficie totale de la ferme				
Moins de 10 acres	3,5		13,2	
10 - 69 acres	15,8		42,3	
70 - 399 acres	62,8		85,0	
400 - 759 acres	78,4		90,5	
760 acres et plus	100,0		100,0	
Valeur totale des produits agricoles vendus				
Moins de \$1,199	7,3		27,7	
\$ 1,200 - \$ 2,499	12,3		41,2	
2,500 - 9,999	29,6		65,4	
10,000 - 49,999	67,8		88,3	
50,000 et plus	100,0		100,0	

<sup>a</sup> Sans Terre-Neuve, le Yukon et les Territoires du Nord-Ouest.  
<sup>b</sup> Les estimations des enquêtes reposent sur un échantillon de 18,327 exploitations.  
<sup>c</sup> Le fichier des enregistrements non apparés comprenait 1,231 exploitations.

En résumé, les résultats de l'analyse des enregistrements non apparés indiquent concrètement que les exploitations omises dans le recensement sont généralement celles dont la production et la valeur agricoles sont au-dessous de la moyenne, hypothèse très répandue mais non prouvée auparavant.

5.4 Analyses des enregistrements apparés

La deuxième série d'analyses portait sur les exploitations agricoles qui ont été dénombrées dans le recensement et les enquêtes et dont les enregistrements ont pu être apparés par ordinateur ou manuellement. Étant donné qu'on a supposé que ces enregistrements du recensement et des enquêtes correspondaient au même ensemble d'exploitations agricoles, on a éliminé les effets qu'aurait pu être attribués à des différences dans le champ d'observation. Pour réduire les effets de l'imputation des données, on a exclu les enregistrements des non-répondants dont toutes les données avaient dû être imputées. Il était donc possible de préciser la nature et l'ampleur des différences entre les réponses fournies au recensement et aux enquêtes. Avant d'examiner les résultats des analyses des enregistrements apparés, on doit toutefois décrire certaines limites inhérentes au fichier des enregistrements apparés, qui ont eu des répercussions sur le processus d'évaluation.

Bien qu'on ait essayé autant que possible de réduire la probabilité d'une erreur dans l'appariement des enregistrements, il se peut qu'un petit nombre des exploitations du fichier des

agricoles et 3,90% de la superficie des terres en culture, mais près de 9,70% du nombre total de fermes. Ces résultats constituaient un premier indice que les "exploitations omises" ne sont pas représentatives de l'ensemble de la population et que leur superficie et d'autres caractéristiques sont en-dessous de la moyenne. Il était donc impossible de mesurer l'ampleur du sous-dénombrement uniquement en fonction du nombre d'exploitations omises. Il fallait plutôt définir les caractéristiques des exploitations omises en examinant les chiffres sur certains produits agricoles.

Pour mieux comprendre la nature du sous-dénombrement, on a ventilé le nombre estimé d'exploitations omises selon les fourchettes de valeurs établies pour la superficie, les ventes, le bétail et d'autres produits. Une comparaison de ces résultats avec ceux calculés à l'échelle de l'univers des enquêtes a révélé que la proportion d'exploitations omises ayant de petites superficies et des chiffres de vente faibles était plus élevée que dans la population visée. Ainsi, on peut constater au tableau 4 qu'une superficie totale de moins de 70 acres a été déclarée par 42,30% du nombre estimé d'exploitations omises, alors que cette proportion ne s'élevait qu'à 15,80% dans l'ensemble de la population cible des enquêtes. Il est estimé que 27,70% des exploitations omises avaient des ventes de moins de \$1,200, contre seulement 7,30% dans l'univers des enquêtes.

Pour comparer ces ventilations, on a également calculé le rapport entre les estimations relatives aux exploitations omises et celles relatives à l'ensemble de la population visée dans les enquêtes, c'est-à-dire le poids des exploitations omises dans les totaux estimés à partir des données des enquêtes. Comme on peut le voir au tableau 4, le fichier des enregistrements non apparés contenait 36,50% du nombre estimé d'exploitations ayant une superficie de moins de 10 acres, la fourchette de valeurs la plus basse établie pour cette variable, mais seulement 4,30% du nombre estimé d'exploitations ayant une superficie de 760 acres ou plus, la catégorie des valeurs les plus élevées. De même, 36,80% du nombre estimé d'exploitations ayant des ventes de moins de \$1,200 figuraient dans le fichier des enregistrements non apparés, mais seulement 3,50% des exploitations ayant des ventes de \$50,000 ou plus étaient dans ce fichier.

Tableau 3

Comparaison des estimations de la superficie et de l'utilisation de la terre dans l'EDA-EDEA pour l'ensemble des fermes et les fermes non apparées (milliers d'acres), 1981, Canada<sup>a</sup>

Variable	Estimation pour l'ensemble du fichier de l'EDA-EDEA <sup>b</sup>	Estimation pour l'EDA-EDEA <sup>c</sup> non apparés de l'ensemble des enregistrements non apparés	Proportion de l'estimation totale dans les enregistrements non apparés
Nombre de fermes	319,476	30,975	9.7
Superficie totale des fermes	175,543	7,768	4.4
Terre améliorée, total	114,610	4,502	3.9
En culture	78,211	2,792	3.6
Pâturage amélioré	9,460	603	6.4
Jachère	24,939	1,004	4.0
Autres terre améliorée	1,999	104	5.2
Terre non améliorée, total	60,933	3,266	5.4
Terre à bois	17,751	1,325	7.5
Autre terre non améliorée	43,182	1,941	4.5

<sup>a</sup> Sans Terre-Neuve, le Yukon et les Territoires du Nord-Ouest.

<sup>b</sup> Les estimations des enquêtes reposent sur un échantillon de 18,327 exploitations.

<sup>c</sup> Le fichier des enregistrements non apparés comprenait 1,231 exploitations.



J. Smith, alors que le fichier des enquêtes en contient un pour J. Smyth; le nom James Smith peut figurer sur un enregistrement et Jim Smith sur un autre, ou encore la ville de St. Catharines peut avoir été introduite sous la forme ("St. Catharines"). Si les données du recensement sur une société en nom collectif ou une corporation ont été fournies par un autre associé ou gestionnaire que celles d'une des enquêtes, il était alors impossible d'appartier par ordinateur les enregistrements correspondants en fonction du nom de l'exploitant. C'est aussi pour cette raison que l'ordinateur ne pouvait pas appartier les enregistrements sur une exploitation agricole si l'exploitant au moment du recensement n'était plus le même au moment des enquêtes.

Pour améliorer le taux d'appartierment et éliminer le biais qui peut entacher le fichier des enregistrements appartés à cause des exploitations qui n'ont pas pu être appartées par ordinateur, une vérification manuelle de ces dernières a été entreprise. À l'aide d'autres données tirées des questionnaires d'enquête comme, par exemple, les raisons sociales ou le nom d'une exploitation, les adresses et les noms des associés, la description des terres de chaque exploitation et des observations inscrites par les interviewers, des employés de soutien ont parcouru les enregistrements non appartés dans le but de trouver la ferme de recensement qui correspondait à chaque enregistrement non apparté. Parmi les 4,459 exploitations du fichier des enquêtes qui n'étaient pas encore retracées, 3,228 ont été retrouvées au cours de la comparaison manuelle. Enfin, 93,3% des 18,327 enregistrements de l'EDA et de l'EDFA pour l'ensemble du Canada ont été appartés aux enregistrements correspondants du fichier du recensement.

Si plus de temps et de ressources avaient pu être consacrés à cette opération, il se peut que les 6,7% d'enregistrements des enquêtes qui n'ont pas été appartés au fichier du recensement aient pu l'être. Toutefois, un grand nombre des enregistrements non appartés des enquêtes ou du recensement ne contenait pas les identifiants nécessaires, et il aurait fallu dépouiller des dossiers administratifs ou entrer en contact avec les exploitants mêmes. Étant donné que les avantages possibles ne semblaient pas en justifier le coût, les recherches manuelles n'ont pas été poursuivies.

Les analyses qui ont pu être faites grâce à l'appartierment se répartissent en deux catégories: celles des enregistrements non appartés des enquêtes et celles des enregistrements appartés du recensement et des enquêtes.

### 5.3 Analyses des enregistrements non appartés

Pour étudier le genre de fermes sous-dénombrées dans le recensement, on a supposé que les enregistrements non appartés représentent bien les exploitations agricoles qui devaient être incluses dans le recensement mais ont été omises. Il était reconnu que les enregistrements non appartés exagèrent l'importance des fermes oubliées à cause de certaines propriétés des sources de données et de l'algorithme d'appartierment. Par exemple, il est probable que des enregistrements dans le fichier des enquêtes avaient un homologue dans le fichier du recensement mais ne pouvaient pas être appartés, que ce soit par ordinateur ou manuellement, parce que le nom ou l'adresse manquait ou était inexact.

Comme il était possible de surestimer les totaux relatifs aux caractéristiques et aux produits des exploitations agricoles omises dans le recensement, il fallait interpréter avec prudence les estimations obtenues à partir des enregistrements non appartés. Malgré tout, les estimations calculées pour l'ensemble du Canada ont été très utiles en tant qu'indicateurs préliminaires des catégories d'exploitations sous-dénombrées dans le recensement.

D'abord, les données de chaque enregistrement ont été multipliées par des facteurs d'expansion pour calculer des estimations relatives aux produits des "exploitations omises". On a ensuite comparé les résultats aux estimations obtenues pour toute la population visée par l'enquête et on a évalué la proportion des estimations totales que représente chaque estimation relative aux "exploitations omises". Cette proportion a été comparée avec le rapport en pourcentage entre le nombre d'exploitations omises et l'estimation du nombre total d'exploitations. Par exemple, le tableau 3 révèle que les exploitations figurant dans le fichier des enregistrements non appartés représentaient seulement 4,4% de l'estimation de la superficie totale des exploitations

Tableau 2  
Comparaison des estimations du recensement et de l'EDA-EDEA  
relatives au nombres de fermes selon la forme  
juridique, 1981, Canada<sup>a</sup>

Forme juridique	Estimation du recensement <sup>b</sup>	Estimation des enquêtes <sup>c</sup>	Écart en pourcentage <sup>d</sup>
Nombre total de fermes	309,410 <sup>e</sup>	319,476	- 3.2 ± 2.6
Ferme individuelle ou familiale	268,199	267,396	0.3 ± 3.0
Société en nom collectif	11,160 <sup>e</sup>	15,908	- 29.8 ± 16.7
- avec convention écrite	17,646 <sup>e</sup>	22,855	- 22.8 ± 10.8
- sans convention écrite	11,744	12,160	3.4 ± 10.4
Corporation	661 <sup>e</sup>	1,142	- 42.1 ± 13.0
Autre genre			

Voir les renvois du tableau 1.

ayant peu de rapport avec l'agriculture ou produire des différences dans la classification des fermes selon la catégorie d'utilisation de la terre. Le micro-couplage offre le mécanisme nécessaire pour étudier ce genre de problème.

## 5.2 Micro-couplage

L'expérience acquise dans d'autres recensements et enquêtes agricoles démontre que même les efforts les plus méticuleux ne peuvent éliminer complètement les erreurs de réponse. Malgré tout le soin apporté à la rédaction de questions claires et non ambiguës, les données recueillies ne peuvent échapper à des problèmes comme les différences dans l'interprétation de certains termes agricoles d'une région à l'autre du Canada ou l'absence d'un accord général sur la manière de classer certaines utilisations de la terre. Les erreurs d'interprétation sont particulièrement fréquentes dans le cas des questions qui ne relèvent pas directement de l'économie ou de l'agriculture ou qui ne s'adressent pas à la plupart des répondants. C'est le micro-couplage, c'est-à-dire l'appariement des enregistrements du recensement avec ceux de l'EDA et de l'EDEA, qui a permis d'évaluer l'effet des erreurs de réponse sur le recensement de l'agriculture de 1981. L'appariement des enregistrements du recensement avec ceux de l'EDA et de l'EDEA a été exécuté en fonction du nom de l'exploitant, de l'adresse, du numéro de téléphone et du code postal de chaque exploitation agricole. Cette opération s'est déroulée en treize étapes qui comportaient la comparaison d'une combinaison différente des identificateurs ou de leurs composantes. À chaque étape, les enregistrements des enquêtes qui n'avaient pas encore été appariés étaient repérés et une recherche informatisée des enregistrements correspondants était exécutée. Pour chaque exploitation agricole visée, les variables ou les clés d'appariement étaient comparées, un caractère à la fois, avec la zone correspondante des enregistrements non appariés du fichier du recensement. Un appariement avait lieu si tous les caractères des clés d'identification étaient les mêmes. Pour l'ensemble du Canada, 75.7% des 18,327 enregistrements des enquêtes ont été appariés. Il était inévitable que, pour un certain nombre d'enregistrements du fichier des enquêtes, aucune ferme correspondante ne puisse être repérée dans le fichier du recensement. Beaucoup d'enregistrements n'ont pas été appariés à cause d'erreurs dans l'orthographe des noms et dans les adresses, qui ont été commises à la collecte ou à la saisie des données du recensement ou des enquêtes. Par exemple, le fichier du recensement peut contenir un enregistrement pour



Tableau 1  
Comparaison des estimations du recensement et de l'EDDA-EDDA  
relatives au nombres de fermes, à l'utilisation de la  
terre (milliers d'acres), 1981, Canada<sup>a</sup>

Variable	Estimation du recensement <sup>b,c</sup>	Estimation des enquêtes <sup>d</sup>	Ecart en pourcentage <sup>d</sup>
Nombre total de fermes	309,410 <sup>e</sup>	319,476	- 3.2 ± 2.6
Superficie totale des fermes	159,866 <sup>e</sup>	175,543	- 8.9 ± 2.4
Terre améliorée	112,390	114,610	- 1.9 ± 2.3
En culture	75,532 <sup>e</sup>	78,211	- 3.4 ± 2.2
Pâturage amélioré	10,523 <sup>e</sup>	9,460	11.2 ± 7.3
Jachère	23,827 <sup>e</sup>	24,939	- 4.5 ± 3.7
Autre terre améliorée	2,509 <sup>e</sup>	1,999	25.5 ± 7.7
Terre non améliorée	47,477 <sup>e</sup>	60,933	- 22.1 ± 4.3
Terre à bois	8,211 <sup>e</sup>	17,751	- 53.7 ± 3.9
Autre terre non améliorée	39,265 <sup>e</sup>	43,182	- 9.1 ± 6.5

<sup>a</sup> Sans Terre-Neuve, le Yukon et les Territoires du Nord-Ouest.

<sup>b</sup> Sans les régions marginales spécifiées et les fermes non incluses dans l'univers des enquêtes. Les totaux du recensement et de l'enquête peuvent ne pas correspondre à la somme des parties à cause de l'arrondissement. Les estimations des enquêtes pour l'ensemble du Canada sont basées sur un échantillon de 18,327 fermes.

<sup>c</sup> Ecart en pourcentage =  $\frac{\text{Estimation du recensement} - \text{Estimation des enquêtes}}{\text{Estimation des enquêtes}} \times 100$ ; l'écart en pourcentage peut ne pas être comparable avec les totaux présentés à cause de l'arrondissement. L'intervalle de confiance indiqué, qui tient compte de l'erreur d'échantillonnage dans les enquêtes, est égal à  $\pm 2 \times$  (coefficient de variation d'une estimation des enquêtes)  $\times$   $\frac{\text{Estimation des enquêtes}}{\text{Estimation du recensement}}$ . Un astérisque signalant un écart significatif figure vis-à-vis des estimations du recensement qui se trouvent à l'extérieur de l'intervalle de confiance de 95% des enquêtes.

différences les plus importantes en ce qui concerne la superficie des fermes se notent toutefois dans les catégories de terre non améliorée, en particulier celle des terres à bois. Ces résultats ont motivé une analyse détaillée de la réponse "terre à bois" qui est résumée à la section 5.5. Les distributions de fréquences estimées à partir des fichiers du recensement et des enquêtes comme certaines fonctions de variables telles que la forme juridique, la superficie totale, la superficie des terres en culture et les ventes ont été comparées. Les différences entre la répartition obtenue dans le recensement et les enquêtes révèlent si des fermes ayant certaines caractéristiques ont pu être omises ou incluses par erreur.

Le tableau 2 présente la ventilation du nombre de fermes estimé selon la forme juridique pour l'ensemble du Canada. Aucune grande différence n'a été notée dans les estimations relatives aux fermes individuelles ou familiales ou aux corporations. Par contre, une analyse détaillée de la couverture des sociétés en nom collectif a été entreprise à cause des grands écarts constatés dans cette catégorie.

L'utilité des macro-comparaisons dans l'évaluation de la couverture est limitée parce qu'on ne peut pas faire abstraction des erreurs de réponse. Par exemple, il semble que l'ampleur et le sens des écarts observés au tableau 1 entre les estimations du recensement et celles des enquêtes dans les catégories de terres améliorées varient trop pour qu'on puisse les attribuer seulement à des erreurs de couverture. Les divergences dans les chiffres concernant les terres à bois découlent peut-être de facteurs autres que la couverture. Il est possible que des différences dans les opérations sur le terrain ou dans la composition des questionnaires puissent causer des écarts entre les résultats du recensement et ceux des enquêtes, faire inclure ou exclure des exploitations

L'évaluation repose sur les données de l'EDA et de l'EDBA menées le 1<sup>er</sup> juillet 1981, environ un mois après le 3 juin, date du recensement. On a supposé que certaines données, comme celles sur les dépenses d'exploitation de l'année précédente, étaient peu déformées par la différence entre les dates de référence. Par contre, il était probable que d'autres réponses puissent avoir changé entre le 3 juin et le 1<sup>er</sup> juillet. Ainsi, l'écart le plus marqué devait paraître dans les données sur le bétail à cause des variations constantes de la taille des troupes aux naissances, aux morts, aux achats, aux transferts, etc. Pour venir à bout de cette difficulté, on a demandé aux exploitants participant aux enquêtes d'indiquer les changements du nombre de bovins et de porcs entre le 3 juin et le 1<sup>er</sup> juillet. Une évaluation a révélé que les données ainsi obtenues ont permis de remédier dans une certaine mesure au problème des différences de référence, mais que le taux de non-réponse était élevé et que la précision des renseignements fournis était douteuse. Par conséquent, il a fallu prendre en considération la différence entre la date de référence du recensement et celle des enquêtes dans la comparaison de toutes les variables susceptibles de changer en fonction de la date des réponses.

Les échantillons de l'EDA et de l'EDBA sont choisis à partir d'une base areolaire de secteurs de dénombrement agricoles et d'une liste des principales exploitations productrices de certains produits importants. Les données sont recueillies par des enquêteurs formés à cet effet au cours d'une interview sur place avec l'exploitant de chacune des fermes choisies. Après les étapes de traitement nécessaires, une méthode d'estimation est utilisée pour étendre les totaux obtenus des échantillons des enquêtes à l'ensemble de la population cible. (Pour de plus amples renseignements sur le plan de sondage, voir Statistique Canada (1984) et Phillips (1978).)

Les estimations des enquêtes présentent les mêmes types d'erreurs non dues à l'échantillon-nage que le recensement. Cependant, étant donné que les enquêtes sont limitées à un nombre moins élevé d'exploitations et que les méthodes de contrôle sont donc meilleures que dans le recensement, il s'ensuit que l'ampleur de ce ces erreurs devrait être moins grande dans les enquêtes que dans le recensement. Les estimations des enquêtes constituaient donc des approximations acceptables des vraies valeurs des données. En revanche, ces estimations renfermaient une erreur d'échantillonnage dont il a fallu tenir compte dans les comparaisons avec les estimations globales obtenues dans le recensement.

## 5.1 Macro-comparaisons

Avant de procéder à l'évaluation au moyen du fichier d'enregistrements appariés, on a examiné les estimations globales calculées à partir des fichiers de données du recensement et des deux enquêtes. Étant donné que l'univers des enquêtes et du recensement était comparable, ces macro-comparaisons à l'échelle des provinces et des régions ont donné une idée préliminaire de la qualité de la couverture du recensement. En comparant les estimations ponctuelles issues du recensement avec la confiance de 95% constituée à partir des enquêtes pour les totaux relatifs au bétail, aux superficies des terres en culture et à d'autres caractéristiques, on a pu détecter des sous-estimations ou des surestimations possibles. On a ensuite poursuivi l'analyse des différences entre les valeurs globales pour voir si ces différences étaient circonscrites dans des catégories particulières de variables. Les analyses globales ont également permis aux personnes qui les faisaient de bien se familiariser avec les deux ensembles de données pour les analyses détaillées effectuées par la suite.

Un exemple des résultats des comparaisons globales est présenté au tableau 1 qui contient les estimations relatives au nombre de fermes, à la superficie totale d'exploitation agricole et à l'utilisation de la terre pour l'ensemble du Canada. Dans la superficie totale d'exploitation agricole, on constate une différence nette entre le chiffre du recensement et l'estimation calculée à partir des données des enquêtes, mais l'ampleur et le sens de cet écart varient beaucoup d'une sous-catégorie à une autre de la variable "utilisation de la terre". La divergence entre les estimations du recensement et celles issues des enquêtes atteint  $25.5\% \pm 7.7\%$  pour la catégorie "autre terre améliorée", mais seulement  $-3.4\% \pm 2.2\%$  pour la catégorie "terre en culture". Les

été recueillies, la méthode de collecte, les notions et les définitions de base et la période de référence. Étant donné que les analyses et les classements recoupés de variables sont plus nuancés dans les évaluations détaillées que dans les évaluations globales, l'effet de ces différences peut être beaucoup plus grand dans les analyses détaillées que dans les analyses globales.

Pour examiner l'utilisation de l'appariement dans l'analyse de l'erreur non due à l'échantillonnage, nous prendrons l'exemple de l'évaluation de la qualité des données dans le recensement de l'agriculture au Canada de 1981. Les analyses globales et détaillées dans cette évaluation ont été faites à l'aide de données recueillies indépendamment dans l'enquête descriptive sur l'agriculture (EDA) et l'enquête sur les exploitations agricoles (EDEA), deux enquêtes probabilitistes de Statistique Canada.

## 5. COMPARAISON DU RECENSEMENT ET DES ENQUÊTES

Le recensement de l'agriculture a eu lieu le 3 juin 1981 dans le cadre des opérations sur le terrain du recensement quinquennal de la population et du logement. Il avait pour objet de recueillir des données sur toutes les fermes de recensement du Canada, soit les fermes, ranchs ou autres exploitations agricoles ayant tiré \$250 ou plus de la vente de produits agricoles au cours des douze mois précédant le jour du recensement ou ayant la capacité de percevoir un montant équivalent au cours des douze mois après le recensement. Au moment de la livraison du questionnaire du recensement de la population et du logement, les recenseurs devaient demander à chaque ménage si un de ses membres avait une ferme ou une autre exploitation agricole correspondant à cette définition. Le cas échéant, les recenseurs remettaient également une formule du recensement de l'agriculture à l'intention de l'exploitant.

Pour améliorer la couverture, on a établi une liste des principales exploitations productrices de certains biens agricoles à partir des résultats du recensement de 1976 et des enquêtes agricoles menées par la suite. Les recenseurs devaient ensuite expliquer s'ils avaient trouvé chacune de ces "fermes spécifiées" dans leur secteur de dénombrement.

Le questionnaire était livré avant le 3 juin à l'exploitant de chaque ferme de recensement et devait être rempli par le répondant lui-même le jour du recensement. Les questions portaient sur les cultures, le bétail, l'utilisation des terres, les ventes, les dépenses et d'autres renseignements utiles aux secteurs public et privé. Des précisions sur la méthodologie et le contenu du recensement de l'agriculture de 1981 sont présentées dans une publication de Statistique Canada (1982).

Le champ d'observation de l'enquête descriptive sur l'agriculture (EDA) et de l'enquête sur les exploitations agricoles (EDEA) inclut la plupart des terres agricoles du Canada. L'EDEA vise le Manitoba, la Saskatchewan et l'Alberta dans les provinces des Prairies et le district de Peace River en Colombie-Britannique, tandis que l'EDA vise l'Île-du-Prince-Édouard, la Nouvelle-Écosse, le Nouveau-Brunswick, le Québec et l'Ontario. Leur univers englobe les exploitations agricoles qui satisfont à la définition d'une ferme de recensement. Toutefois, les exploitations ayant peu d'importance économique, comme les fermes exploitées par des institutions et celles situées dans des régions comme les centres urbains, où l'activité agricole est faible ou nulle, sont exclues. Les exploitations non comprises dans la population visée par ces enquêtes ont été supprimées du fichier du recensement pour rendre les univers comparables. Les enregistrements supprimés représentaient seulement 2,8% des fermes et 1,8% de la superficie totale d'exploitation agricole calculée à partir de l'ensemble du fichier du recensement.

Ces deux enquêtes probabilitistes recueillent des données sur les mêmes grandes variables agricoles que le recensement, c'est-à-dire les cultures, le bétail, l'utilisation des terres, les dépenses d'exploitation et ainsi de suite, et reposent sur les mêmes notions et définitions. Il existe toutefois des différences dans le libellé et la structure de certaines questions et dans les instructions concernant ce qu'il faut inclure ou exclure de certaines réponses. Tel qu'il est mentionné plus bas dans la description des résultats, il a fallu tenir compte de l'effet de ces différences dans la comparaison des données.



les données de la Current Population Survey aux États-Unis ont été appariées à celles du recensement de la population (U.S. Bureau of the Census 1964) et, au Canada de nouveau, le fichier de l'enquête descriptive sur l'agriculture a été apparié à celui du recensement de l'agriculture (Statistique Canada 1979).

Pour réduire le fardeau des répondants, des données administratives sont utilisées de plus en plus dans les évaluations. Andersen et coll. (1979) et Horvitz (1981) ont apparié les dossiers de médecins et d'hôpitaux aux résultats d'enquêtes sur l'état de santé. Des dossiers relatifs à l'immigration et aux naissances ont été utilisés au cours de l'opération de contre-vérification des dossiers dans le recensement de la population au Canada (Krotki 1980) et, aux États-Unis, les dossiers fiscaux de l'IRS ont été utilisés pour analyser les erreurs de réponse dans le recensement de la population (U.S. Bureau of the Census 1970). D'autres appariements avec des données administratives ont été faits par Foulkenberry et Tortora (1981) dans l'évaluation des résultats d'une enquête agricole et par Marquis, Marquis et Polich (1981) dans une analyse des réponses à des questions délicates.

En général, la qualité d'une évaluation fondée sur l'appariement d'enregistrements dépend de deux facteurs:

- 1) la qualité des données de la source de comparaison à laquelle les données d'une enquête sont appariées et,
- 2) le degré d'univocité des identificateurs utilisés pour l'appariement et la précision de la technique d'appariement proprement dite.

Tout d'abord, par la définition même des objectifs d'une évaluation, les données de comparaison doivent servir d'approximations des vraies valeurs. Des erreurs aléatoires dans les données qui s'annulent dans l'ensemble des enregistrements n'ont pas d'incidence marquée sur les comparaisons globales. En revanche, ce genre d'erreur peut avoir des effets graves sur les résultats d'analyses concernant des enregistrements particuliers.

On peut supposer que certaines bases de données choisies comme repères pour des comparaisons ne contiennent pas d'erreurs. Les dossiers des naissances, par exemple, offrent des renseignements très précis sur le lieu de naissance et l'âge. On sait parfois que certains ensembles de données renferment des erreurs de réponse ou de couverture mais, si ces erreurs sont mesurables, on peut les prendre en considération dans l'analyse et compter de supposer qu'il n'y a pas d'erreur. Par contre, dans bien des cas, les données de comparaison contiennent des erreurs qu'il est impossible de cerner ou de mesurer avec exactitude. Les meilleures approximations des vraies valeurs proviennent alors de l'ensemble de données qui comporte le moins d'erreurs. Des données qui ont été recueillies à l'aide de méthodes plus exactes ou par un personnel mieux qualifié que dans le cas de l'enquête soumise à l'évaluation peuvent être considérées comme ayant moins d'erreurs et comme étant une base raisonnable pour établir des comparaisons.

Le deuxième grand facteur qui détermine la qualité d'une évaluation tient à l'opération d'appariement proprement dite. Il arrive qu'un numéro d'identification unique ait été attribué à chaque membre de la population et que ce numéro soit fidèlement reproduit dans les deux fichiers de données qu'on veut comparer. L'appariement peut alors se faire directement grâce au caractère univoque de ce genre d'identificateur. En revanche, il est impossible que les meilleurs identificateurs dans certains cas soient des caractéristiques non univoques telles que des noms, qui peuvent renfermer des erreurs introduites lors de la collecte ou de la saisie des données.

L'algorithme d'appariement peut influencer la qualité d'une évaluation, surtout quand les clés d'appariement ou les identificateurs ne sont pas parfaits. Cet algorithme risque d'appariement ou d'empêcher l'appariement d'enregistrements correspondants quand les valeurs clés sont différentes, à cause d'une petite erreur ou d'un oubli. La fréquence possible de ce genre d'erreur peut avoir un effet important sur la composition des fichiers des enregistrements appariés et non appariés.

Parmi les autres facteurs qui ont des effets sur la comparabilité de deux ensembles de données, même dans les évaluations globales, il y a les différences dans la date où les données ont

de la Current Population Survey à ceux du recensement de la population, tandis que les statisticiens au Canada peuvent comparer les données de l'enquête sur la population active à celles du recensement de la population ou les chiffres de l'enquête descriptive sur l'agriculture avec ceux du recensement de l'agriculture (Statistique Canada 1979). Des projections démographiques ont également été utilisées à titre d'approximations des valeurs exactes, comme, par exemple, dans les comparaisons de l'enquête sur la population active et du recensement de la population du Canada décrites par Fellegi (1973).

Les données administratives sont de plus en plus utilisées dans les évaluations à cause de la préoccupation grandissante concernant le fardeau imposé aux répondants d'enquêtes. Les estimations produites à partir de dossiers comme ceux des organismes responsables des impôts sur le revenu, des allocations familiales, de l'immatriculation des véhicules automobiles et de la commercialisation de produits agricoles peuvent servir d'approximations des vraies valeurs recherchées. Par exemple, Statistique Canada examine actuellement la possibilité d'employer les dossiers fiscaux pour recueillir et évaluer les données sur le revenu agricole.

Les quelques exemples mentionnés plus haut montrent qu'un grand éventail de sources de données ont été utilisées pour comparer les estimations calculées à partir d'une enquête. Des comparaisons globales donnent une idée de l'erreur totale qui entache les résultats d'une enquête, toutefois, elles ne permettent pas de préciser ou de mesurer les composantes de cet erreur. Les comparaisons globales ne nous renseignent pas sur les effets des erreurs de couverture et de réponse. On doit donc utiliser d'autres méthodes qui facilitent l'analyse détaillée de l'erreur totale observée.

Il existe une technique qui peut fournir beaucoup d'informations sur les erreurs et leurs causes: l'appariement des enregistrements contenant les données obtenues dans une enquête et des enregistrements d'un fichier de données comparables. Les enregistrements appariés permettent de comparer les valeurs recueillies dans une enquête à celles de la source repère, qu'on suppose exactes. À l'aide de classements recoupés en fonction de différentes variables, on peut étudier les caractéristiques des unités dont certains résultats ne correspondent pas parfaitement. En examinant les enregistrements qui ne peuvent pas être appariés, on peut également déceler les unités susceptibles d'avoir été incluses ou exclues par erreur de couverture de l'enquête.

Dans cette étude, nous nous penchons sur l'utilisation de l'appariement pour l'évaluation de l'erreur non due à l'échantillonnage. On décrit les avantages et les lacunes de cette méthode et les types d'analyses qu'elle rend possible. Pour illustrer l'application de la technique d'appariement, on résume la méthodologie et quelques résultats de l'évaluation de la qualité des données du recensement de l'agriculture au Canada de 1981.

#### 4. UTILISATION DE L'APPARIEMENT

Un examen sommaire des études de l'évaluation de l'erreur totale dans les enquêtes révèle un grand nombre d'analyses basées sur la méthode d'appariement. Comme dans le cas des comparaisons globales, une grande variété de sources de données comparables ont été utilisées. Dans un bon nombre de cas, des enquêtes ont été menées postérieurement à une autre enquête afin de recueillir des données qui soient de meilleure qualité et plus détaillées. Par exemple, des experts ont procédé à des réentrevues dans une étude sur le prix marchand des maisons (Kish et Lansing 1954); un suivi a eu lieu pour évaluer le recensement de l'agriculture aux États-Unis (U.S. Bureau of the Census 1982); l'enquête sur la population active au Canada comprend un programme de réentrevues (Tremblay, Singh et Clavel 1976) et Statistique Canada effectue une vérification des logements vacants après le recensement de la population et du logement (Statistique Canada 1980).

Les fichiers de recensements et d'enquêtes exécutés indépendamment ont aussi été appariés pour évaluer la qualité des données. Par exemple, les résultats de l'enquête sur la population active au Canada ont déjà été appariés à ceux du recensement de la population (Krotki 1980);

Des modèles ont par la suite été conçus pour évaluer les effets d'erreurs non dues à l'échantillonnage indépendamment et en corrélation avec d'autres erreurs. À titre d'exemples, mentionnons le modèle d'erreur du U.S. Bureau of the Census résumé par Nisselson et Bailar (1976) et coll. (1979), lequel repose sur un modèle décrit quelques années auparavant par Kish (1965) et coll. (1979). Hartley (1981) construit un modèle comportant des termes pour les erreurs attribuables aux interviewers, aux codeurs et aux répondants et propose un plan d'échantillonnage qui facilite l'estimation de ces erreurs.

Divers auteurs classent les composantes de l'erreur non due à l'échantillonnage en différentes catégories. Dans la présente étude, nous répartissons les erreurs non dues à l'échantillonnage en deux groupes: les erreurs de couverture et les erreurs de réponse. Une erreur de couverture se produit quand une unité appartenant à l'univers visé est oubliée ou comptée plus d'une fois ou quand une unité n'appartenant pas à cet univers est incluse. À cause des erreurs de couverture, toutes les données des unités mal dénombrées sont soit incluses, soit exclues quand c'est le contraire qui est prévu. Ces erreurs peuvent donc altérer certaines ou toutes les estimations. Les erreurs de réponse faussent des estimations particulières calculées à partir des données sur les unités incluses à juste titre dans l'échantillon. Ces erreurs peuvent s'introduire au moment de la collecte des données ou au cours des étapes de traitement subséquentes. Parmi les sources d'erreur possibles, il y a les erreurs d'interprétation des questions commises par le répondant, la non-réponse partielle ou totale, l'influence de l'interviewer et les erreurs commises pendant la saisie ou le codage des données.

### 3. ÉVALUATION DE L'ERREUR NON DUE À L'ÉCHANTILLONNAGE

Dans la plupart des modèles d'erreurs dans les enquêtes, l'erreur est définie comme étant l'écart entre la valeur exacte d'une variable et l'estimation calculée à partir d'un échantillon. Théoriquement, pour évaluer l'erreur non due à l'échantillonnage, on doit donc comparer les résultats d'une enquête et la vraie valeur d'une variable d'intérêt. En pratique, toutefois, il est rare qu'on connaisse la valeur exacte d'un caractère et il peut même être impossible de la connaître. On est donc obligé de comparer les données d'enquête à celles d'une autre source qui offre les approximations les plus rapprochées des valeurs exactes.

Quand on choisit la source de données qui représente le mieux les vraies valeurs inconnues, il faut tenir compte de certains facteurs. Les données de la source de comparaison doivent avoir été recueillies indépendamment de celles de l'enquête soumises à une évaluation. Idéalement, les définitions et les notions employées dans la collecte des données, de même que les périodes de référence de ces données doivent être identiques pour chacune des deux sources. L'univers, des données de comparaison doit aussi être le même que dans l'enquête évaluée ou, du moins, on doit pouvoir extraire des sous-univers comparables. On doit également disposer de tous les renseignements sur l'objectif de la source de comparaison, les méthodes de collecte des données et les traitements ou les mises à jour effectuées par la suite. Enfin, ce qui est peut-être le plus important, la qualité des données de référence doit être assez élevée pour permettre une comparaison critique des données de l'enquête qu'on évalue.

Dans la réalité, bien entendu, il est rare qu'une seule source de référence puisse satisfaire tous ces critères. Il arrive qu'une source soit excellente pour une sous-population en particulier, mais qu'une autre source soit meilleure pour le reste de la population. On peut parfois modifier les données pour supprimer les différences entre les dates de référence ou les définitions des variables dans les deux ensembles qu'on veut comparer. Toutefois, même si on possède les meilleures estimations des valeurs exactes, la plupart des sources de données comportent des lacunes importantes, et il peut être impossible de mesurer, voire de décrire, comment elles perturbent la comparaison. On peut obtenir des approximations des valeurs exactes à partir de diverses sources. On peut comparer les estimations ou d'une enquête à celles d'un autre recensement ou d'une autre enquête. Ainsi, aux États-Unis, les statisticiens peuvent comparer les résultats



# Appariement d'enregistrements pour l'évaluation des erreurs non dues à l'échantillonnage dans le recensement de l'agriculture de 1981 au Canada

J. COULTER<sup>1</sup>

## RÉSUMÉ

Cette étude décrit l'utilisation de l'appariement de fichiers de données comparables dans l'évaluation de l'erreur non due à l'échantillonnage. Pour illustrer cette technique, on explique comment la qualité des données du recensement de l'agriculture de 1981 au Canada a été évaluée et on présente quelques résultats de cette analyse.

**MOTS CLÉS:** Erreur non due à l'échantillonnage; Couverture; Erreur de réponse; Appariement; Appariement des enregistrements; Recensement de l'agriculture.

## 1. INTRODUCTION

Tout au long de l'évolution des méthodes d'échantillonnage probabilistes pour la collecte de données, l'évaluation et le contrôle des erreurs d'échantillonnage sont demeurés des préoccupations constantes. Beaucoup de recherches ont été faites pour mettre au point des plans d'échantillonnage permettant de réduire l'erreur d'échantillonnage et de la mesurer facilement. Toutefois, une grande partie de l'erreur totale dans beaucoup d'enquêtes est attribuable non à la technique d'échantillonnage, mais aux effets d'autres aspects du processus de collecte des données. Dans les recensements, en particulier, où on recueille des données auprès de 100% de la population cible, l'erreur d'échantillonnage est nulle. Les erreurs proviennent plutôt des répondants, des interviewers, des codeurs, des préposés à l'entrée des données et d'autres personnes au cours des étapes de la collecte, de la saisie et du traitement. Plus notre compréhension de l'effet des erreurs non dues à l'échantillonnage sur la qualité des données augmente, plus l'élaboration des moyens de contrôle et de mesure prend de l'importance.

## 2. MODÈLES DE L'ERREUR DANS LES ENQUÊTES

Les premiers travaux sur l'erreur totale dans les enquêtes, comme celui de Deming (1944), résument les sources d'erreur possibles et décrivent la nécessité de prendre en compte leurs divers effets dans la planification des opérations de collecte de données. Quand l'étude de l'erreur dans les enquêtes a pris de l'ampleur, des auteurs tels que Hansen et coll. (1951), Sukhatme et Seth (1952), Hansen, Hurwitz et Bershad (1961) et d'autres ont proposé des modèles généraux pour décrire les composantes de l'erreur d'échantillonnage et de l'erreur non due à l'échantillonnage. Des études ont été faites sur les corrélations entre les erreurs imputables aux interviewers ou aux codeurs, et des méthodes ont été mises au point pour mesurer les répercussions de ces erreurs. Fellegi (1964) a présenté un modèle détaillé qui tient compte des corrélations entre les erreurs provenant de beaucoup de sources différentes.

J. Coulter, Division des opérations du recensement, Statistique Canada, 2<sup>e</sup> étage, Édifice Jean Talon, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Cholette: Estimateurs des cycles économiques

## 7. HISTORIQUE DES MOYENNES PAR MINIMISATION QUADRATIQUE

Cette approche de minimisation quadratique provient de Whittaker (1923). Leser (1961 et 1963) montra comment la minimisation quadratique pouvait servir à mettre au point des moyennes mobiles cyclique. Cholette (1980) proposa des substituts aux moyennes mobiles deux de douze et deux de quatre. Ces substituts furent incorporés dans la méthode de désaisonnalisation X-11-ARMMI de Dagum (1980) à titre optionnel.

La moyenne cyclique semestrielle présentée dans ce travail pourrait aussi être incorporée dans une méthode du genre X-11 (ou X-11-ARMMI). Ceci permettrait la désaisonnalisation des séries semestrielles et le calcul de leurs facteurs saisonniers grâce aux moyennes mobiles saisonnières couramment utilisées pour les séries mensuelles et trimestrielles.

## 8. CONCLUSION

Ce travail a présenté une moyenne mobile à cinq termes qui élimine la saisonnalité des séries semestrielles. L'estimateur reproduit plus fidèlement les cycles économiques que la moyenne mobile deux de deux. La deux de deux a aussi le désavantage de ne pas fournir d'estimation pour les premier et dernier semestres de la série.

## BIBLIOGRAPHIE

AKAIKE, H. and ISHIGURO, M. (1980). *Bayesian Seasonal Adjustment Program*. The Institute of Statistical Mathematics, Computer Science Monograph No. 13, Tokyo.

BOX, J.E.P., and JENKINS, G.M. (1970). *Time Series Analysis, Forecasting and Control*. Holden-Day.

CHOLETTE, P.A. (1980). A Comparison of Various Trend-Cycle Estimators. *Time Series Analysis*, (O.D. Anderson and M.R. Perryman Eds.), 77-87.

DAGUM, E.B. (1980). *La méthode de désaisonnalisation X-11-ARMMI*. Car. 12-564F, Statistique Canada.

KOOPMANS, L.H. (1974). *The Spectral Analysis of Time Series*. Academic Press.

LAROCQUE, G. (1977). Analyse d'une méthode de désaisonnalisation: le programme X-11 du U.S. Bureau of the Census, version trimestrielle. *Annales de l'I.N.S.E.E.*, 28, 105-127.

LESER, C.E.V. (1961). A Simple Method of Trend Construction. *Journal of the Royal Statistical Society Series B*, Vol. 23, 91-107.

LESER, C.E.V. (1963). Estimation of Quasi-Linear Trend and Seasonal Variation. *Journal of the American Statistical Association*, 58, 1033-1043.

MACAULY, F.R. (1931). *The Smoothing of Time Series*. National Bureau of Economic Research, Washington.

PHILIPS, L., and BLOMIE, R. (1973). *Analyse chronologique*. (Vander Eds.). Louvain, 334.

SCHLICHT, E. (1981). A Seasonal Adjustment Principle and a Seasonal Adjustment Method Derived from this Principle. *Journal of the American Statistical Association*, 76, 374-378.

SHISKIN, J., YOUNG, A.H. and MUSGRAVE, J.C. (1967). *The X-11 Variant of Census the Method II Seasonal Adjustment Program*. Technical Paper No. 15, U.S. Bureau of the Census.

WHITTAKER, E. (1923). On a New Method of Graduation. *Proceedings of the Edinburgh Mathematical Society*, 41, 63-75.

Si l'on omettait les estimations obtenues au moyen de la dernière pondération, plusieurs faux signaux *disparaissent*, mais au prix de l'actualité des estimations. Ceci illustre le dilemme du statisticien entre l'actualité et la fiabilité des estimations, quelle que soit la méthode d'estimation. (En pratique, l'analyste sérieux attend au moins une confirmation d'un signal avant d'y croire.) A part les cinq faux signaux mentionnés, les estimations préliminaires affichent un mouvement très semblable et parfois indiscernable de celui des estimations finales.

## 6. ETABLISSEMENT DES POIDS DE LA MOYENNE CYCLIQUE SEMESTRIELLE

La série observée  $z_t$  comprend la tendance-cycle  $c_t$  à estimer et un résidu saisonnier-irrégulier  $s_t + e_t$  ( $= z_t - c_t$ ):

$$(1) \quad z_t = c_t + (s_t + e_t), \quad t = 1, \dots, 5.$$

Suivant l'approche de Leser (1961 et 1963) et de Cholette (1980), la tendance-cycle recherchée minimise la somme quadratique de quadruples différences (premier terme de (2)). Sur l'intervalle d'estimation de cinq semestres, la composante s'approxime donc autant que possible à un polynôme temporel du troisième degré. Cette spécification permet l'estimation sur l'intervalle d'un cycle économique complet avec ses quatre phases d'expansion, de retournement, de récession et de reprise. Les résidus saisonniers-irréguliers ( $z_t - c_t$ ) minimisent la somme quadratique des premières différences saisonnières sur semestres homologues (deuxième terme de (2)). Cette spécification signifie que le résidu saisonnier-irrégulier d'un semestre doit ressembler le plus possible au résidu correspondant au même semestre de l'année voisine.

Les résidus saisonniers-irréguliers minimisent en outre la somme quadratique de leurs sommes sur deux semestres consécutifs (troisième terme de (2)). Ce critère indique que la saisonnalité de deux semestres successifs devrait s'annuler, que l'irrégularité de deux semestres successifs devrait s'annuler et que l'irrégularité ne devrait pas affecter le niveau de la tendance-cycle recherchée.

Ces trois critères spécifiés pour les composantes se combinent en la fonction objective suivante:

$$(2) \quad f(c) = \sum_{t=1}^5 (c_t - 4c_{t-1} + 6c_{t-2} - 4c_{t-3} + c_{t-4})^2 + \sum_{t=2}^5 \{(z_t - c_t) + (z_{t-1} - c_{t-1})\}^2$$

On peut reprendre l'équation (3) en algèbre linéaire:

$$(3) \quad F(C) = C'A'AC + (Z - C)'B'B(Z - C) + (Z - C)'F'F(Z - C) = C'HC + (C - Z)'G(C - Z)$$

où A, B et F désignent respectivement les opérateurs matriciels de quadruples différences, de premières saisonnières et de sommes annuelles définies de la façon suivante:

$$A = \begin{bmatrix} 1 & -4 & 6 & -4 & 1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 \\ 1 & 0 & -1 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \quad F = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Les équations normales associées à (3) s'écrivent

$$(4) \quad dF/dC = 2HC + 2G(C - Z) = 0$$

et impliquent la solution:

$$(5) \quad C = (H + G)^{-1}GZ = WZ.$$

5. ANALYSE GRAPHIQUE DES ESTIMATIONS TERMINALES

On trouve dans la figure 4 les estimations préliminaires obtenues à l'aide des deux pondérations terminales pour les années 1968 à 1980, accompagnées des estimations centrales finales disponibles correspondantes. La figure 4 a) montre les estimations terminales tombant dans le deuxième semestre; et la figure 4 b), dans le premier semestre. (Un seul graphique aurait été trop congestionné.) Si on considère les estimations centrales comme vraies (ou du moins comme plus fiables), il apparaît que les estimations terminales provoquent cinq faux signaux: en 1968 (flèche dans la fig. 4 b), en 1972 (4 a), en 1974 (4 b), en 1975 (4 a) et en 1976 (4 b). Un faux signal est réputé survenir ici lorsque les estimations terminales indiquent un changement de direction de la tendance-cycle et que ce changement se trouve contredit plus tard par les estimations centrales finales devenant disponibles (grâce à de nouvelles observations). Ces faux signaux tendent à se manifester lorsque la série ralentit son mouvement dans une direction et reprend sa course dans la même direction. Lorsqu'il y a un changement marqué de direction comme en 1978, ceci ne semble pas se produire.

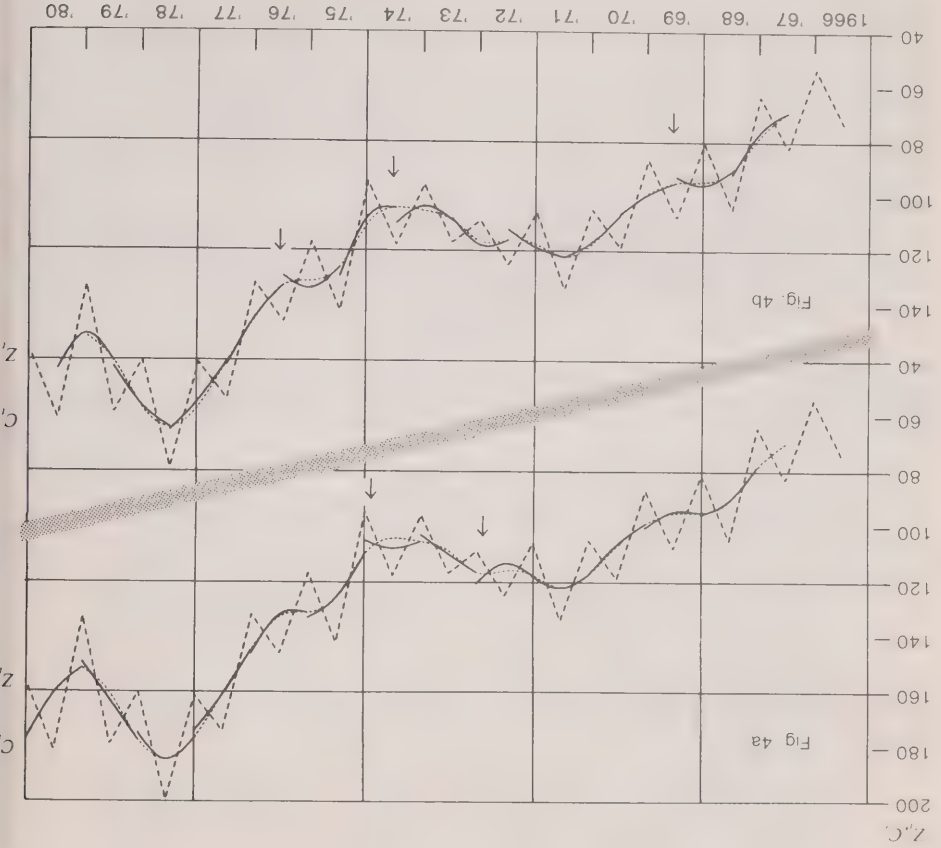


Figure 4. Série semestrielle avec saisonnalité (----); estimations préliminaires de sa tendance-cycle par les poids terminaux (—) de la moyenne mobile cyclique semestrielle proposée a) pour les seconds semestres et b) pour les premiers semestres; estimations finales (....) par les poids centraux de la moyenne.

Fig. 3

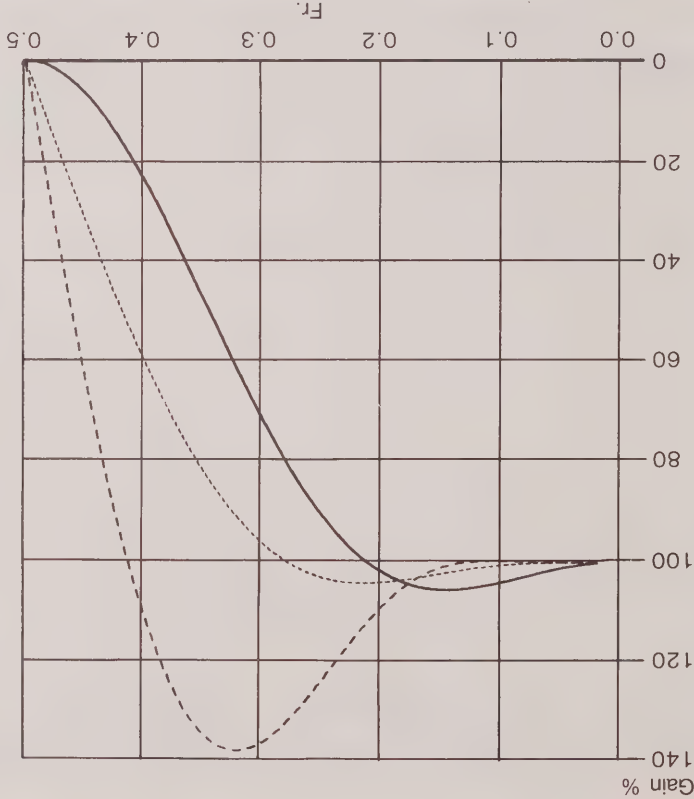


Figure 3. Fonctions de gain de la pondération centrale modifiée (—) et des pondérations relatives à l'avant-dernier (...) et au dernier (---) estimés

Tableau 2

Déphasages enregistrés par les pondérations terminales à certaines fréquences d'intérêt en nombre de semestres

		avant - dernier		dernier	
fréquences cycliques:		estimé		estimé	
0.100 (10 semestres)	0.01	0.01	0.01	0.01	0.01
0.167 ( 6 semestres)	0.05	0.05	0.05	0.05	0.05
0.200 ( 5 semestres)	0.09	0.09	0.16	0.00	0.05
0.250 ( 4 semestres)	0.16	0.16	0.28	0.17	0.00
0.333 ( 3 semestres)	0.28	0.28	0.46	0.45	0.17
fréquences saisonnières					
0.467	0.46	0.46	0.48	0.47	0.45
0.483	0.50	0.50	0.50	0.50	0.483
0.500 ( 2 semestres)	0.50	0.50	0.50	0.50	0.500 ( 2 semestres)



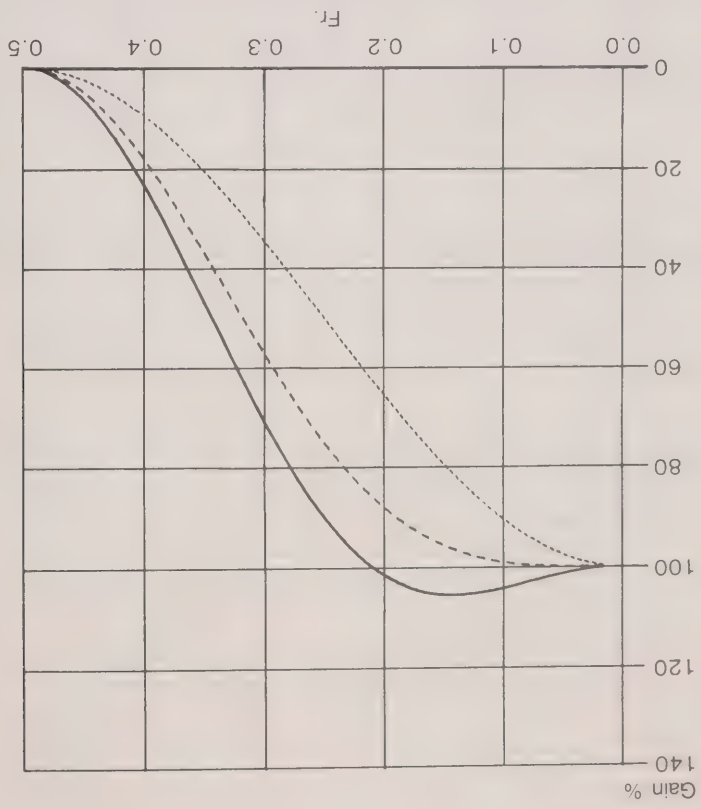


Figure 2. Fonctions de gain des pondérations centrales modifiées (—) et non-modifiées (----) de la moyenne cyclique semestrielle proposée et de la moyenne mobile *deux de deux* (....)

et 0,483; et, encore plus des fréquence aléatoires. Du point de vue du gain, la pondération relative à l'avant-dernier estimé devrait donner des estimés moins fiables que la pondération centrale modifiée. La situation se dégrade davantage pour la pondération relative au dernier estimé (ligne en tirets de la figure 3). On observe une forte amplification de certaines fréquences aléatoires et cycliques rapides (gains atteignant parfois 137%). Il faut par conséquent user de prudence dans l'interprétation de l'estimé résultant de ces poids. Il y aurait peut-être lieu de rejeter le dernier estimé complètement pour les séries répétées irrégulières (contenant ces fréquences amplifiées).

Comme on peut le voir dans le tableau 1, les pondérations terminales ne sont pas symétriques. Elles engendrent par conséquent des déphasages, consignés en nombre de semestres dans le tableau 2, pour certaines fréquences d'intérêt. Aux fréquences cycliques cibles, le déphasage s'avère modérément faible pour l'avant-dernière pondération. Dans ce dernier cas, une ondulation cyclique de cinq semestres accusera un retard de 0,09 semestre dans les estimés; une de quatre semestres, de 0,16 semestre; et une de trois semestres, de 0,28 semestre; etc.

Le déphasage atteint son maximum à la fréquence saisonnière fondamentale (0,500). Mais ceci n'a aucune importance pour celle-ci, puisque la pondération l'élimine complètement. Le déphasage importe un peu pour les fréquences de saisonnalité mobile 0,467 et 0,483, qui ne sont pas complètement éliminées.



4. ANALYSE SPECTRALE DE LA MOYENNE

Les courbes des figures 2 et 3 représentent les fonctions de gain de la pondération étudiée. La valeur du gain en ordonnée indique en pourcentage la fraction des onduations sinusoïdales conservées, c'est-à-dire *passées* aux estimés par la pondération (Laroque, 1977; Wallis, 1982). La fréquence des onduations apparaît en abscisse et varie de 0 à 0.500. La fréquence 0.500 correspond à une onduation annuelle de deux semestres (1/50), c'est-à-dire à la saisonnalité stable. La saisonnalité mobile est prise en compte par les quelques fréquences quasi-annuelles voisines: 0.467 et 0.483. La fréquence 0.333 correspond à une onduation de trois semestres (1/33); la fréquence 0.250, quatre semestres; 0.200, cinq semestres; 0.167, six semestres, etc. Les fréquences associées à des onduations de trois semestres et plus (à gauche de 0.333 dans les figures 2 et 3) appartiennent à la tendance-cycle des séries et constituent les fréquences cibles de l'estimateur.

Les fréquences comprises entre 0.333 et 0.467 exclusivement sont associées à des fluctuations de périodicités inférieures à un an et demi et supérieures aux périodicités saisonnières quasi-annuelles. Elles correspondent à la composante aléatoire des séries. Une moyenne cyclique idéale devrait éliminer 100% de ces fréquences irrégulières, éliminer 100% des fréquences saisonnières et quasi saisonnières et ne conserver que les fréquences cycliques 0 à 0.333 inclusivement.

1) analyse de poids centraux

La courbe continue de la figure 2 montre que la pondération centrale modifiée de la moyenne cyclique semestrielle conserve 100% de toutes les onduations de 5 semestres (2 ans et demi) et plus. En effet, la courbe réside au-dessus de 100% partout à gauche de la fréquence 0.200. En comparaison, la moyenne mobile *deux de deux*, représentée par la courbe pointillée, conserve seulement 65% des onduations de 5 semestres et 93% de celles de 10 semestres. La pondération centrale modifiée passe en outre 55% des onduations de 3 semestres et 90% de celles de semestres (2 ans), contre 25 et 50% respectivement pour la *deux de deux*.

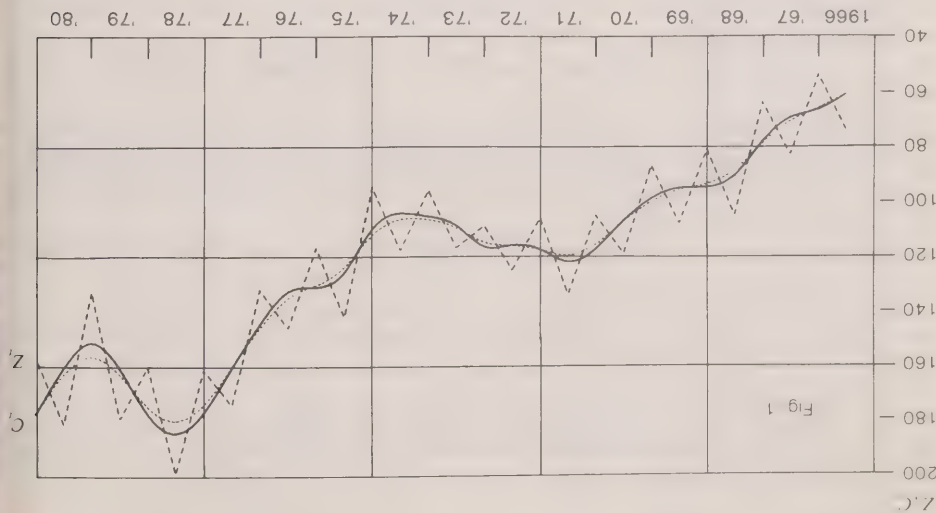
Les deux pondérations éliminent complètement la saisonnalité stable avec des gains valant zéro à la fréquence saisonnière 0.500; ainsi que pratiquement toute la saisonnalité mobile. Par contre, la *deux de deux* élimine un peu plus des fréquences irrégulières que la pondération centrale modifiée. Dans le choix entre les deux moyennes, on fait face au dilemme suivant: conférer aux estimations d'avantage de mouvements cycliques mais aussi d'avantage d'irrégularité ou bien moins de mouvements cycliques (surtout les plus rapides) que d'irrégularité, la pondération centrale modifiée de la moyenne mobile proposée est certainement plus indiquée que la *deux de deux*.

La courbe en tirets de la figure 2 représente le gain de la pondération centrale non modifiée de la moyenne cyclique semestrielle, telle qu'obtenue dans la section 6. Aux fréquences cycliques, sa performance est supérieure à celle de la *deux de deux*; mais, inférieure à celle de la pondération centrale modifiée. Par exemple cette dernière reproduit 101% des onduations de 5 semestres (fréquences 0.200) contre 88% pour la non modifiée. Quant à l'amplification de 5% (gain de 105%) enregistrée par une réduction comparable de 6% (gains de 94%) avec la pondération non modifiée. En effet, l'analyse des poids centraux

Idéalement, les gains des pondérations terminales devraient être identiques au gain de la pondération centrale. Dans pareil cas, les poids terminaux et centraux auraient le même effet sur la série traitée (sauf pour des possibles déphasages).

Le gain de la pondération relative à l'avant-dernier estimé (courbe pointillée de la figure 3) est assez sensible au gain de la pondération centrale modifiée (courbe continue). À remarquer que le premier ressemble davantage au dernier qu'au gain de la pondération centrale non modifiée de la figure 2. C'est pourquoi nous avons modifié cette pondération.

La pondération associée à l'avant-dernier estimé conserve les fréquences cycliques et élimine la saisonnalité stable. Cependant, elle préserve entre 11 et 21% des fréquences de saisonnalité mobile 0.467



### 3. POIDS DE LA MOYENNE

Le tableau 1-A consigne les valeurs exactes des poids de la moyenne mobile cyclique des périodes 3 à 28 (dans la figure 1); la deuxième, les poids terminaux servant à l'avant-dernier estimé; et la troisième, au dernier estimé. Le tableau 1-B montre les poids centraux trouvés selon la méthodologie exposée à la section 6. Cependant, nous avons jugé bon de les remplacer par les poids centraux *modifiés* du tableau 1-A pour des raisons à élucider plus bas.

Tableau 1-A  
Poids exacts de la moyenne cyclique semestrielle proposée

pondération centrale modifiée	0.1000	0.2500	0.7000	0.2500	0.7500	0.0625	0.9375
pondération avant-dernière	0.0625	0.2500	0.3750	0.3750	0.7500	0.0625	
pondération dernière							

Tableau 1-B  
Pondération centrale non modifiée

	0.0625	0.2500	0.6250	0.2500	0.0625
--	--------	--------	--------	--------	--------

# Estimateurs des cycles économiques dans les séries semestrielles

PIERRE A. CHOLETTE<sup>1</sup>

## RÉSUMÉ

Ce travail offre une moyenne mobile qui élimine la saisonnalité et élimine la saisonnalité des séries semestrielles (observables deux fois l'an). La moyenne proposée conserve en entier la puissance des cycles de trois ans et plus; 90% de ceux de deux ans; et 55% des cycles d'un an et demi. Par comparaison, la moyenne mobile *deux de deux* retient respectivement la puissance de 75%, 50% et 25% des mêmes cycles.

**MOTS CLÉS:** moyenne mobiles; cycles économiques; analyse spectrale.

## 1. INTRODUCTION

Il existe apparemment des séries semestrielles auxquelles ne correspond aucune donnée mensuelle. Dans pareil cas, l'obtention des chiffres désaisonnalisés semestriels à partir de valeurs mensuelles désaisonnalisées s'avère impossible. À notre connaissance, il n'existe aucune méthode de désaisonnalisation pour les séries semestrielles non plus.

Ce travail présente une moyenne mobile qui élimine la saisonnalité et calcule la tendance cycle des séries semestrielles. L'approche de minimisation quadratique empruntée remonte à Whittaker (1923) et fut ensuite reprise par Leser (1961 et 1963), Cholette (1980), Schlicht (1981) et par d'autres. La moyenne obtenue a cinq termes. Elle comprend une pondération centrale pour les semestres centraux des séries; et deux pondérations terminales pour les deux premiers et derniers semestres. Par conséquent, il n'y a donc pas de perte d'observation aux extrémités des séries, comme avec la moyenne mobile *deux de quatre* (utilisée pour les séries trimestrielles) par exemple.

Les propriétés spectrales de la pondération centrale s'avèrent supérieures à celles de la moyenne mobile *deux de deux*, qui vient d'abord à l'esprit pour traiter des séries semestrielles. Les propriétés des pondérations terminales seront aussi examinées.

## 2. ILLUSTRATION DE LA MOYENNE

La figure 1 montre la série semestrielle originale observée  $z_t$  (en tirets) de 1966 à 1980, accompagnée de la tendance-cycle  $c_t$  (ligne continue) estimée à l'aide de la moyenne cyclique semestrielle présentée dans ce travail. Comme espéré, la tendance-cycle se comporte de manière lisse et manifeste des cycles assez rapides, notamment un cycle de trois ans, allant du deuxième semestre de 1977 au premier semestre de 1980. On trouve une estimation pour chaque observation, y compris les deux premières et dernières observations. La tendance-cycle produite par la moyenne mobile *deux de deux* (courbe pointillée) ne donne pas d'estimation pour le premier et le dernier semestres de la série. En outre la *deux de deux* escamote les creux et les sommets cycliques en comparaison de la moyenne proposée.

<sup>1</sup> Pierre A. Cholette, Séries chronologiques recherche et analyse, Statistique Canada, 25<sup>e</sup> étage, Édifice R.H. Coats, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

- HIDIROGLU, M.A. et RAO, J.N.K. (1983). Chi-square tests for the analysis of three-way contingency tables from the Canada Health Survey. Technical Report, Statistics Canada.
- JAMREY, P.B., KOCH, G.G. et STOKES (1981). Categorical data analysis: Some reflections of the log linear model and logistic regression. Part I: Historical Methodological Overview. *International Statistical Review*, 49, 265-283.
- JOHNSON, N.L. et KOTZ, S. (1970). Continuous Univariate Distributions. Boston: Houghton Mifflin.
- KOCH, G.G., FREEMAN, D.H. JR., and FREEMAN, J.L. (1975). "Strategies in the Multivariate Analysis of Data From Complex Surveys". *International Statistical Review*, 43, 59-78.
- RAO, J.N.K. et SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-square tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- RAO, J.N.K. et SCOTT, A.J. (1983). On Chi-square tests for multiway contingency tables with cell proportions estimates from survey data. Carleton Mathematical Series No. 199, Carleton University, Ottawa.
- SATFERTHWAIT, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.
- SHAH, B.V. (1981). Development of survey data analysis software. Research Triangle Institute, Research Triangle Park, North Carolina.
- SHUSTER, J.J. et DOWNING, D.J. (1976). Two-way contingency tables for complex sampling schemes. *Biometrika*, 63, 272-276.



pas des variables indépendantes. Cette possibilité est particulièrement importante quand  $X_j$  est une variable (comme la strate géographique) qui est utilisée dans la structure du plan d'échantillonnage. Étant donné que ce genre de variable est habituellement connue pour toutes les unités, on peut soit (a) ne plus accorder d'importance à la question de l'indépendance, soit (b) mettre  $X_j$  en marge et s'intéresser seulement à  $X_1$  et  $X_2$ , non à leur distribution conditionnelle. Si on choisit (b), les résultats présentés dans les sections précédentes semblent appropriés. Dans certains cas, il peut être possible de vérifier si  $X_1$  et  $X_2$  sont conditionnellement indépendants pour  $Y_j$  donné.

Toutefois, il existe une autre difficulté. Supposons que nous voulons connaître les proportions  $\pi_{ij}$  dans chaque case d'un tableau pour une population finie de taille  $N$ . Dans un recensement de cette population, il est peu probable qu'on obtienne la relation  $\pi_{ij} = \pi_{i+} \pi_{+j}$  exactement. Au mieux, on peut souhaiter qu'une mesure quelconque d'association comme, par exemple  $N\Delta(\pi_{ij} - \pi_{i+} \pi_{+j})^2 / \pi_{i+} \pi_{+j}$ , soit faible. Notons que même dans un modèle de superpopulation à indépendance exacte, on ne s'attendrait pas à ce que ce coefficient d'association soit nul. Il faut peut-être en définitive tester des hypothèses telles que

$$H_0: \text{coefficient d'association} < C$$

$$H_1: \text{coefficient d'association} > C.$$

D'autres recherches doivent être menées sur ce sujet. Toutefois, dans des cas pratiques où la traction de sondage n'est pas grande, les méthodes décrites ici sont applicables.

## BIBLIOGRAPHIE

- ALTHAM, P.A.E. (1976). "Discrete Variable Analysis for Individuals Grouped Into Families", *Biometrika*, 63, 263-269.
- BINDER, D.A. (1983). Les variances d'estimateurs asymptotiquement normaux basés sur des enquêtes complexes. Version française d'un article qui sera publié dans *International Statistical Review*, 51.
- BISHOP, Y.M.M., FIENBERG, S.E. et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press.
- BRIER, S.S. (1978). "Discrete Data Models With Random Effects", Technical Report, University of Minnesota, School of Statistics.
- COHEN, J.E. (1976). "The Distribution of the Chi-Squared Statistic Under Cluster Sampling Form Contingency Tables", *Journal of the American Statistical Association*, 71, 665-670.
- COWAN, J. et BINDER, D.A. (1978). The effect of a two-stage sample design on tests of independence in a 2 by 2 table. *Techniques d'enquête*, 4, 16-28.
- FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of American Statistical Association*, 80, 148-157.
- FELLEGG, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 71, 665-670.
- FIENBERG, S.E. (1980). *The Analysis of Cross Classified Data*, (2<sup>e</sup> ed.). Cambridge, Mass: MIT Press.
- GRIZZLE, J.E., STARMER, C.F. et KOCH, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, 25, 489-504.
- HIDIROGLOU, M.A., FULLER, W.A. et HICKMAN, R.D. (1980). MINICARP: A program for estimating simple descriptive statistics and their variances for multi-stage stratified designs. Iowa State University: Ames, Iowa.
- HIDIROGLOU, M.A. et RAO, J.N.K. (1981). Tests chi-carré pour l'analyse d'observations classées de l'Enquête Santé Canada. Version française d'un document présenté aux réunions de l'Institut international de statistique, Buenos Aires, 1981.



A l'heure actuelle, il n'existe pas de logiciel intégré comme ceux mentionnés plus haut qui soit conçu pour des analyses des types de données dont il est question dans les sections précédentes. Par conséquent, quiconque cherche une solution rapide à un problème de ce genre doit habituellement avoir recours aux logiciels existants, lesquels risquent de ne pas être appropriés.

Les possibilités qui s'offrent sont les suivantes:

- utiliser les logiciels existants, mais en les modifiant;
- écrire des programmes spéciaux qui existent déjà;
- adopter une combinaison des choix ci-dessus.

Pour les analyses présentées dans ce document, c'est en modifiant le programme MINI CARP (Hidiroglou, Fuller et Hickman, 1980) qu'on a obtenu les résultats des exemples 1, 2 et 3. Pour l'exemple 4, on a utilisé une combinaison de programmes écrits en PL/1 et en SAS. L'analyse des données de l'Enquête sur la Population Active (exemple 5) a été faite à l'aide de SAS et de programmes sur mesure.

Chacune des possibilités énumérées plus haut présente des inconvénients sur le plan pratique, notamment:

- a) s'il est nécessaire de modifier un logiciel qui existe déjà, il faut posséder une connaissance détaillée de ses mécanismes;
- b) il peut être nécessaire de reproduire des renseignements identiques sur différents fichiers de données, étant donné que les solutions susmentionnées ne peuvent pas être intégrées comme les systèmes généraux;
- c) en comparaison des logiciels intégrés qui sont conçus de manière à être facilement utilisables par différents utilisateurs, les autres programmes manquent souvent d'élégance et d'efficacité opérationnelle;
- d) il est souvent impossible d'obtenir une documentation complète sur les programmes spéciaux ou faits sur mesure, ce qui limite l'accessibilité de ces logiciels.

Ds travaux sont actuellement en cours pour mettre au point des programmes écrits en SAS qui permettent d'effectuer un grand nombre des analyses décrites ici. Notre objectif final est semblable à celui formulé par Shah (1981), c'est-à-dire élaborer un logiciel complet pour l'analyse des données d'enquête. Tous les efforts visant à atteindre cet objectif sont justifiés si nous voulons éviter les problèmes qu'éprouvent actuellement les spécialistes qui doivent soit construire leurs propres programmes, soit utiliser les logiciels qui existent déjà, ce qui peut produire des résultats et des conclusions erronés.

## 7. CONCLUSION

Nous avons examiné quelques-uns des problèmes que pose l'ajustement de modèles à des données qualitatives recueillies en fonction de plans d'échantillonnage complexes. La méthode fondamentale qu'on utilise ici est de calculer la statistique de Wald qui convient au modèle ajusté ou d'employer le test approprié sur les plans fondés sur l'échantillonnage multinomial et en suite trouver une bonne approximation de la distribution sous l'hypothèse nulle.

Nous n'avons pas abordé la question du choix entre les méthodes basées sur le modèle et celles basées sur le plan de sondage. On a plutôt mis l'accent sur le processus d'inférence basé sur le plan de sondage.

Pour résumer cette dichotomie, prenons de nouveau le test d'indépendance dans un tableau de contingence à deux dimensions. L'indépendance nous intéresse quand nous voulons savoir si la valeur de la variable  $X_1$  a un effet sur notre connaissance de la variable  $X_2$ . Si la réponse est non pour toutes les unités de la population, on dit alors que ces variables sont indépendantes. Mais si nous connaissons la valeur de  $X_1$ , il est encore possible que  $X_1$  et  $X_2$  ne soient

où  $V'' = V - \{p_j - \hat{f}_j\}$ . On peut calculer les  $\{V''_i\}$  en utilisant la relation  $\bar{p} - \bar{f} = [I - \text{diag}\{f'_i(1-f'_i)\}\hat{A}(\hat{A}'\hat{A})^{-1}\hat{A}'\bar{D}N](\bar{p} - \bar{f})$ .

### Exemple 5

Les données de l'enquête sur la population active d'octobre 1980 ont été utilisées pour ajuster des modèles logistiques (logit) de la probabilité d'avoir un travail. L'échantillon est composé d'hommes âgés de 15 à 64 ans qui font partie de la population active et n'étudient pas à plein temps. Un modèle logit, quadratique pour les variables âge et instruction, a été ajusté à ces données. On a défini des groupes d'âge en divisant l'intervalle [15, 64] en dix parties et le milieu de chaque groupe d'âge est représenté par l'intervalle  $[10 + 5j, 14 + 5j]$ ,  $j = 1, 2, \dots, 10$ . Le milieu de chaque groupe d'âge a été utilisé comme la valeur de l'âge de toutes les personnes appartenant au groupe correspondant. On a établi six niveaux d'instruction en attribuant à chaque personne une valeur basée sur la médiane du nombre d'années d'instruction. Ainsi, la création de catégories d'âge et d'instruction produit un tableau comprenant soixante cases.

Soit  $\pi_i = \Pr\{\text{une personne classée dans la } i^{\text{ème}} \text{ case a un emploi}\}$ ,  $i = 1, 2, \dots, 60$ . Nous supposons que  $0 < \pi_i < 1$ . Par conséquent,  $1 - \pi_i$  représente la probabilité qu'une personne dans la  $i^{\text{ème}}$  case soit chômeur. Le modèle ajusté à la forme suivante:

$$(1) \quad \ln \frac{1 - \pi_i}{\pi_i} = \beta_0 + \beta_1 a_i + \beta_2 a_i^2 + \beta_3 d_i + \beta_4 d_i^2$$

$$i = 1, 2, \dots, 60$$

où les  $a_i$  et les  $d_i$  sont les valeurs des variables âge et instruction des personnes classées dans la  $i^{\text{ème}}$  case.

À partir des estimations  $p_i$  obtenues pour les  $\pi_i$  dans l'échantillon, les valeurs de la statistique de Pearson,  $W_{11}$ , et du rapport de vraisemblance  $LR_5$  ont été calculées et les résultats sont que  $W_{11} = 98,94$  et  $LR_5 = 101,20$ . La valeur critique de la distribution khi-carré à 55 degrés de liberté qui délimite une région de 5% à droite est 73,31. Les valeurs calculées de  $W_{11}$  et de  $LR_5$  permettent de rejeter le modèle (1). Toutefois, les résultats obtenus pour  $W_{11}$  ou  $LR_5$  sont appropriés seulement si l'échantillon est aléatoire.

L'estimation de la valeur propre moyenne,  $\Sigma \delta/55$ , pour tester la qualité de l'ajustement est égale à 1,88. Ce calcul a pour effet de diminuer  $W_{11}$  à 52,63 et  $LR_5$  à 53,83. Cette correction permet donc de constater que les données concordent avec le modèle (1). On a également songé à employer la statistique de Wald,  $(p - \hat{f})' [V''_0]^{-1} (p - \hat{f})$ , pour tester la qualité de l'ajustement. (Notons que l'inverse généralisé de  $V''_0$  est utilisé pour ce cas puisque cette matrice est singulière.) Quand  $p_j = 1$ , il faut que la valeur de  $p_j$  soit un peu perturbée pour qu'on puisse calculer la statistique de Wald. La statistique de Wald s'avère instable pour notre problème. De faibles perturbations dans les estimations de  $p$  entraînent des changements considérables dans la valeur de la statistique de Wald.

En outre, cette valeur est très élevée ici à cause de l'instabilité de la matrice de variances-covariances estimée qui entre dans le calcul de la statistique de Wald. Cette statistique est au moins trente fois plus élevée que nos valeurs corrigées du khi-carré.

## 6. PROBLÈMES ACTUELS DE L'UTILISATION DE PROJICIELS

Grâce aux progrès réalisés dans la technologie informatique, les opérations de collecte, de stockage et d'accès aux données sont devenues faciles et efficaces. Des systèmes puissants d'application générale, comme TPL, STATAK et ESTIMATION SYSTEM, permettent à des utilisateurs et à des analystes d'estimer des totaux et leur variance avec assez peu de difficulté. En outre, un certain nombre de PROJICIELS analytiques offerts tels BMDP, SPSS et SAS se prêtent extrêmement bien à certains contextes. Mais la capacité de ces PROJICIELS à exécuter des analyses comme celles décrites ici est limitée. Par exemple, dans les tests d'hypothèses ou l'inférence statistique, ces PROJICIELS supposent que les données analysées proviennent d'enquêtes basées sur des échantillons aléatoires simples.

Exemple 4

On a ajusté un modèle de régression logistique à des données de l'Enquête Santé Canada sur 20,726 répondants pour expliquer le fait qu'ils ont ou n'ont pas consulté un médecin sur une période de douze mois. En tout, on a estimé que 77% de la population avait consulté un médecin au moins une fois. Les résultats sont résumés au tableau 3 (pour une description détaillée, voir Binder (1983)). Le modèle de régression logistique a semblé très bien s'ajuster aux données.

5.3 Variables explicatives qualitatives

La théorie décrite dans cette section a été présentée par G. Roberts dans un document non publié (Université Carleton). Dans le cas qui nous occupe, toutes les variables explicatives sont qualitatives. Les domaines sont notés  $\{1, \dots, I\}$ . Nous définissons  $p_i$  l'estimation calculée à partir des données de l'enquête de la proportion du  $i^{\text{ème}}$  domaine et  $N_i$  l'estimation de la taille du  $i^{\text{ème}}$  domaine,  $N_i$ . Suivant le modèle, la proportion espérée dans le  $i^{\text{ème}}$  domaine est  $f_i$  où

$$\log \{f_i / (I - f_i)\} = \bar{q}_i / \theta,$$

pour  $q_i$  connu et  $\theta$  au paramètre inconnu. Nous définissons  $\hat{q} = [\hat{q}_1, \dots, \hat{q}_I]^T$  et  $\hat{D}_N = \text{diag} \{N_1, \dots, N_I\}$ . Selon le modèle, l'estimateur calculé à partir des données de l'enquête de  $f = (f_1, \dots, f_I)^T$  est  $\tilde{f}$ , la solution de l'équation

$$\hat{q}^T \hat{D}_N (p - \tilde{f}) = \bar{q}, \tag{5.1}$$

Puisque asymptotiquement

$$\hat{\theta} - \theta = (\hat{q}^T \Delta \hat{q})^{-1} \hat{q}^T \hat{D}_N (p - \tilde{f}),$$

où  $\hat{\Delta} = \text{diag} \{N_1 f_1 (1 - f_1), \dots, N_I f_I (1 - f_I)\}$ , il s'en suit que

$$n^{1/2}(\hat{\theta} - \theta) \rightarrow N[0, (\hat{q}^T \hat{\Delta} \hat{q})^{-1} \hat{q}^T \hat{D}_N \hat{q} (\hat{q}^T \Delta \hat{q})^{-1}]$$

lorsque  $n^{1/2}(\bar{p} - f) \rightarrow N(0, \hat{D}_p)$ .

Suivant un échantillonnage binomial indépendant, la matrice de variances-covariances est réduite à  $(N/n)(\hat{q}^T \hat{\Delta} \hat{q})^{-1}$ , où  $n$  est la taille de l'échantillon. Le test du rapport de vraisemblance pour tester la qualité de l'ajustement est

$$LR_s = 2(n/N) \sum_{i=1}^I N_i [p_i \log(p_i / f_i) + (1 - p_i) \log((1 - p_i) / (1 - f_i))],$$

où  $n$  est la taille d'échantillon et  $N = \sum N_i$ . Si  $H_0$  est vraie, cette statistique est asymptotiquement équivalente à

$$W_{11} = (n/N) \sum_{i=1}^I N_i (p_i - f_i)^2 / [f_i (1 - f_i)].$$

En général, la distribution de  $LR_s$  sera celle de  $\sum \delta_i Z_i^2$ , où  $\{Z_i\}$  est un ensemble de variables aléatoires indépendantes distribuées selon une loi  $\chi^2_{\delta_i}$ , et  $\{\delta_i\}$  sont les valeurs propres de  $N^{-1} \hat{D}_N [\hat{\Delta}^{-1} - \hat{q}(\hat{q}^T \hat{\Delta} \hat{q})^{-1} \hat{q}^T] \hat{D}_N \hat{q} \hat{D}_N^{-1}$ . En prenant l'espérance mathématique de  $W_{11}$  et l'approximation

$$W_{11} \approx \frac{I}{\sum \delta_i} \chi^2_I,$$

où  $s = \text{rang}(\hat{q})$ , il s'en suit que

$$\sum \delta_i = (n/N) \sum_{i=1}^I N_i v_{ij}'' / \{f_i' (1 - f_i')\}$$

Si  $\bar{X}\bar{\theta} = \bar{X}_1\bar{\theta}_1 + \bar{X}_2\bar{\theta}_2$  et que nous voulons vérifier les hypothèses suivantes:

$$H_0: \bar{\theta}_2 = 0$$

$$H_1: \bar{\theta}_2 \neq 0,$$

on obtient une statistique de Wald qui se définit ainsi:

$$W_{10} = n \bar{\theta}_2^T (\bar{X}_2^T \bar{A} \bar{X}_2)^{-1} \bar{\theta}_2$$

où

$$\bar{X}_2 = [I - \bar{X}(\bar{X}_1^T \bar{A} \bar{X}_1)^{-1} \bar{X}_1^T \bar{A}] \bar{X}_2.$$

Le test du rapport de vraisemblance repose alors sur la statistique

$$LR_4 = 2 \sum_{i=1}^n \left[ y_i \log \left( \frac{n_i}{n} \right) + (1 - y_i) \log \left( \frac{(1 - \frac{n_i}{n})}{(1 - \frac{n}{n})} \right) \right]$$

qui est asymptotiquement équivalente à  $W_{10}$  si  $H_0$  est vraie.

### 5.2 Autres méthodes d'échantillonnage

Supposons maintenant que  $n^{-1/2} \bar{X}_T(\bar{y} - \bar{\pi}) \rightarrow N(0, \bar{V})$  et que  $\bar{V}$  est un estimateur convergent de  $\bar{V}$ . Ici,  $\bar{y}$  n'est pas nécessairement un vecteur composé de 0 et de 1, mais il peut en réalité dépendre des poids d'échantillonnage et d'autres facteurs de correction. On peut normalement estimer que  $\bar{V}$  est une somme d'observations aléatoires et la plupart des plans d'échantillonnage admettent un estimateur convergent d'une somme d'observations (qui ne sont pas nécessairement indépendantes). Pour estimer  $\bar{V}$ , nous utilisons  $\hat{\bar{V}}$  au lieu de  $\bar{\pi}$  dans l'estimateur. Etant donné que, asymptotiquement,

$$(\hat{\theta} - \bar{\theta}) = (\bar{X}_T^T \bar{A} \bar{X}_T)^{-1} \bar{X}_T^T (\bar{y} - \bar{\pi}),$$

il est possible de déduire que

$$n^{1/2} (\hat{\theta} - \bar{\theta}) \rightarrow N(0, n^2 (\bar{X}_T^T \bar{A} \bar{X}_T)^{-1} \bar{X}_T^T \bar{A} \bar{X}_T)^T;$$

voir Binder (1983) pour une explication détaillée de ce résultat. On peut alors construire une statistique de Wald à partir de la matrice de variances-covariances estimée de  $\hat{\theta}_2$ .

Tableau 3

Modèle de régression logistique pour expliquer la consultation d'un médecin

Variable	Type de données	d.l. Statistique de Wald
Age	Qualitatif	4
Sexe	Qualitatif	1
Interaction âge-sexe	Qualitatif	4
Revenu de la famille	Qualitatif	5
Profession	Qualitatif	3
Interaction profession-sexe	Qualitatif	3
Etat matrimonial	Qualitatif	3
Antécédents médicaux	Qualitatif	2
Nombre de problèmes de santé	Qualitatif	1
Consommation de médicaments	Qualitatif	2
Nombre d'accidents	Qualitatif	2
Nombre de jours de maladie	Qualitatif	2
Taille de l'agglomération	Qualitatif	2
Rapport du nombre de médecins et de la population dans la province	Quantitatif	1
		0,540



Exemple 3

Hidiroglou et Rao (1983) ont examiné toutes les estimations directes dans un tableau à trois dimensions de données de l'Enquête Santé Canada recoupées selon la consommation de médicaments (cinq catégories: 0, 1, 2, 3, 4 + classes de médicaments sur une période de deux jours)  $\times$  l'âge (quatre catégories: 0-14, 15-44, 45-64, 65 +)  $\times$  et le sexe (hommes, femmes). Nous résumons ici le résultat d'un test d'indépendance de l'âge et du sexe dans chaque classe de médicament ( $n = 31,668$ ). L'hypothèse nulle est donc

$$H_0: \pi_{ijk} = \pi_{i+} \pi_{j+} \pi_{k+}.$$

Si on utilise la notation de Bishop, Fienberg et Holland (1975), dans laquelle  $\log \pi_{ijk} = n + n^{(i)} + n^{(j)} + n^{(k)} + n^{(ij)} + n^{(ik)} + n^{(jk)} + n^{(ijk)}$ , l'hypothèse nulle peut s'écrire

$$H_0: n^{23(jk)} = n^{123(jk)} = 0 \text{ pour tous les } (i, j, k).$$

La valeur brute de la statistique khi-carré est 23 pour 15 degrés de liberté. La valeur propre moyenne est de 1,39, de sorte que l'approximation réduit la valeur de khi-carré à 16. La valeur non corrigée du khi-carré conduirait donc l'analyste à rejeter l'hypothèse nulle au seuil de 10%, alors que l'approximation indique que  $h_{ij}$  ne peut pas être rejetée, même au seuil de 30%.

5. MODELES DE RÉGRESSION LOGISTIQUE

5.1 Échantillonnage multinomial

Nous examinons maintenant un modèle de régression logistique pour la distribution conditionnelle d'une variable binaire  $y$  en fonction d'un vecteur donné  $\tilde{x}$  de variables indépendantes. Plus précisément, il s'agit de la distribution conditionnelle suivante:

$$\Pr(y_i | x_i) = \pi(\tilde{x}_i)^{y_i} [1 - \pi(\tilde{x}_i)]^{1-y_i},$$

où  $y_i \in \{0, 1\}$ .

Dans le modèle de régression logistique

$$\log \left\{ \frac{\pi(\tilde{x}_i)}{1 - \pi(\tilde{x}_i)} \right\} = \tilde{x}_i^T \tilde{\theta},$$

où  $\tilde{\theta}$  est un vecteur inconnu de paramètres.

Notons que si  $x_i$  est un vecteur qualitatif dont les éléments prennent la valeur 0 ou 1, ce modèle est un cas spécial du modèle log-linéaire décrit à la section 4. On permet ici à  $x_i$  d'être arbitraire. L'extension de cette analyse au cas où la variable  $y$  comprend  $k$  catégories peut se faire directement. Il est également possible d'écrire le modèle de régression logistique sous une forme générale:

$$\log \left\{ \frac{\pi(\tilde{x}_i)}{\pi(\tilde{x}_i^T \tilde{\theta})} \right\} = \mathcal{F}(\tilde{x}_i^T \tilde{\theta}),$$

pour une fonction connue,  $\mathcal{F}(\cdot)$ , mais cet aspect n'est pas abordé ici.

L'estimateur du maximum de vraisemblance de  $\tilde{\theta}$  se calcule à l'aide de l'équation

$$\tilde{0} = \tilde{\tilde{\theta}}$$

où  $\tilde{y} = (y_1, \dots, y_n)^T$ ,  $\tilde{\pi} = [\pi(\tilde{x}_1), \dots, \pi(\tilde{x}_n)]^T$  et  $\tilde{X} = [\tilde{x}_1, \dots, \tilde{x}_n]^T$ . Dans certaines conditions de régularité appropriées,

$$n^{1/2}(\tilde{\theta} - \tilde{\theta}) \rightarrow N[0, n(\tilde{X}^T \tilde{V} \tilde{X})^{-1}], \text{ où } \tilde{V} = - \tilde{D}^* \tilde{D}^* (\tilde{D}^*).$$



de sorte que la statistique de Wald est

$$W_8 = n\hat{\theta}_2^T \hat{X}_2^T \hat{P} \hat{X}_2 \hat{\theta}_2.$$

Si  $H_0$  est vraie, cette statistique est asymptotiquement équivalente à la statistique du khi-carré de Pearson

$$n(\hat{\pi} - \frac{n}{N})^T \hat{D}^{-1}(\hat{\pi} - \frac{n}{N}),$$

où à la statistique du test du rapport de vraisemblance

$$LR_3 = 2n \sum_{i=1}^k p_i \log(\hat{\pi}/\pi).$$

Si  $H_0$  est vraie, ces variables suivent asymptotiquement une loi de type  $\chi^2$ .

#### 4.2 Autres méthodes d'échantillonnage

Nous supposons de nouveau que les proportions dans chaque case,  $\pi$ , satisfont à la condition  $\pi = \log \pi = u(\hat{\theta}_1, \hat{\theta}_2) \hat{1} + \hat{X}_1^T \hat{\theta}_1 + \hat{X}_2^T \hat{\theta}_2$  mais on a maintenant que  $n(\hat{d} - \pi) \rightarrow N(\hat{0}, \bar{V})$ , où  $\bar{d}$  est une estimation calculée à partir des données d'une enquête.

Rao et Scott (1983) ont proposé une statistique de Wald pour tester  $\hat{\theta}_2 = \hat{0}$ . Soit  $\hat{C}$  une matrice de dimension  $k \times s$  pour laquelle  $\hat{C}^T \hat{X}_1 = \hat{0}$ ,  $\hat{C}^T \hat{1} = \hat{0}$  and  $\hat{C}^T \hat{X}_2$  est non singulière. Par exemple, si  $\hat{X}_1^T \hat{X}_2 = \hat{0}$ ,  $\hat{C} = \hat{X}_2$  est une définition convenable. Cela revient à supposer que  $\hat{C}^T \bar{\pi} = \hat{0}$ . Ainsi

$$\hat{C}^T(\bar{\pi} - \pi) = \hat{C}^T \bar{D}^{-1}(\bar{\pi} - \pi)$$

$$= \hat{C}^T \bar{X}(\bar{X}^T \bar{P} \bar{X})^{-1} \bar{X}^T(\bar{d} - \pi),$$

où  $\bar{\pi}$  est reporté de l'équation (4.1), dans laquelle  $\bar{d}$  est une estimation calculée à partir des données d'une enquête.

On a donc la statistique de Wald suivante:

$$W_9 = n\bar{\pi}^T \hat{C}[\hat{C}^T \bar{X}(\bar{X}^T \bar{P} \bar{X})^{-1} \bar{X}^T \bar{P} \bar{X}(\bar{X}^T \bar{P} \bar{X})^{-1} \bar{X}^T \hat{C}]^{-1} \bar{X}^T \hat{C} \bar{d}.$$

Bindar (1983) a obtenu des résultats semblables. Si, sous  $H_1$ , le modèle est saturé ( $r+s = k-1$ ), alors  $\bar{d} = \bar{\pi}$  et on peut écrire

$$W_9 = n\bar{\pi}^T \hat{C}[\hat{C}^T \bar{D}^{-1} \bar{D}^{-1} \hat{C}]^{-1} \hat{C}^T \bar{d}.$$

Rao et Scott (1984) ont démontré que, si on utilise  $\bar{P}$  au lieu de  $\bar{P}$  pour calculer  $W_9$  et  $W_9$ , ces statistiques sont asymptotiquement équivalentes à la statistique du rapport de vraisemblance ou à la statistique du khi-carré de Pearson. Ces auteurs ont aussi montré que la statistique du test du rapport de vraisemblance a une distribution comme celle de  $\sum_{i=1}^s \delta_i Z_i^2$  sous  $H_0$ , où  $\{Z_i^2\}$  est un ensemble de variables indépendantes de type  $\chi_1^2$  et les  $\{\delta_i\}$  sont les valeurs propres de

$$(4.3) \quad (\hat{X}_1^T \hat{P} \hat{X}_1)^{-1} (\hat{X}_1^T \hat{P} \hat{X}_2),$$

où  $\hat{X}_2$  est définie dans l'équation (4.2).

#### 4.3 Approximations

Comme dans les sections précédentes, nous utilisons l'approximation suivante de la distribution sous l'hypothèse nulle

$$\sum_{s=1}^s \delta_i Z_i^2 \approx \left( \frac{\sum_{s=1}^s \delta_i}{s} \right) \chi_s^2.$$

Pour évaluer cette expression, il faut calculer la trace de la matrice (4.3). Rao et Scott (1984) ont démontré que, si le modèle admet des solutions explicites de  $\hat{\pi}$  and  $\hat{\pi}$ , cette approximation dépend de la matrice  $\hat{V}$  seulement en fonction de l'effet du plan de sondage sur les fréquences des cases et les fréquences marginales. Cette observation est particulièrement utile quand on dispose seulement d'estimations de ces effets, ce qui est souvent le cas dans les tableaux publiés.

Le vecteur  $\pi = (\pi_1, \dots, \pi_r)'$  contient les proportions dans chaque case et  $\sum_{j=1}^r \pi_j = 1$ . Nous observons  $n = (n_1, \dots, n_r)'$ , le vecteur des totaux dans chaque case, et à partir d'un échantillon aléatoire, de sorte que  $\hat{n}$  suit une loi multinomiale ( $\sum n_j = n$ ). Définissons  $\hat{p} = \hat{n}/n$  et

$$\hat{\pi} = \log \bar{\pi}.$$

Dans le modèle log-linéaire, le vecteur des paramètres  $\theta = (\theta_1, \dots, \theta_r)'$ , est tel que

$$\hat{\pi}(\theta) = n(\bar{\theta}) \bar{1} + \bar{X}\bar{\theta},$$

où  $\bar{X}$  est une matrice connue de dimension  $k \times l$  et de rang complet et  $\bar{X}'\bar{1} = \bar{0}$ . Notons que  $l \leq k-1$ . Si  $l = k-1$ , le modèle est saturé.

Pour obtenir l'estimateur du maximum de vraisemblance pour  $\theta$ , on doit résoudre l'équation (4.1):

$$X'(\bar{p} - \hat{\pi}) = \bar{0}, \tag{4.1}$$

où  $\hat{\pi} = \bar{\pi}(\bar{\theta})$ . Asymptotiquement,

$$\hat{\pi} - \bar{\pi} \doteq \bar{P}\bar{X}(\bar{\theta} - \bar{\theta}),$$

où  $\bar{P} = \bar{D} - \bar{\pi}\bar{\pi}'$ . La formule (4.1) permet d'écrire

$$\bar{\theta} - \bar{\theta} \doteq (\bar{X}'\bar{P}\bar{X})^{-1}\bar{X}'\bar{D}(\bar{p} - \bar{\pi})$$

et

$$\hat{\pi} - \bar{\pi} \doteq \bar{P}\bar{X}(\bar{X}'\bar{P}\bar{X})^{-1}\bar{X}'\bar{D}(\bar{p} - \bar{\pi}).$$

Étant donné que  $n^{1/2}(\bar{p} - \bar{\pi}) \rightarrow N(\bar{0}, \bar{P})$  on peut écrire

$$n^{1/2}(\bar{\theta} - \bar{\theta}) \rightarrow N[\bar{0}, (\bar{X}'\bar{P}\bar{X})^{-1}]$$

$$n^{1/2}(\bar{\pi} - \bar{\pi}) \rightarrow N[\bar{0}, \bar{P}\bar{X}(\bar{X}'\bar{P}\bar{X})^{-1}\bar{X}'\bar{P}].$$

Supposons maintenant que l'expression linéaire  $\bar{X}\bar{\theta}$  admet une décomposition de la forme  $X_1\bar{\theta}_1 + \bar{X}_2\bar{\theta}_2$  où  $\bar{X}_1$  et  $\bar{X}_2$  sont de rang complet,  $\bar{X}_1$  est de dimension  $k \times r$ ,  $\bar{X}_2$  de dimension  $k \times s$ ,  $\bar{\theta}_1$  de dimension  $r \times 1$  et  $\bar{\theta}_2$  de dimension  $s \times 1$ , où  $r + s = l$ . Nous voulons vérifier l'hypothèse

$$H_0: \bar{\theta}_2 = \bar{0},$$

contre l'hypothèse

$$H_1: \bar{\theta}_2 \neq \bar{0}.$$

Nous utilisons  $\bar{\theta}_1, \bar{\theta}_2, \bar{\pi}$ , etc. pour représenter les estimations calculées à partir du modèle, si  $H_1$  est vraie, et  $\bar{\theta}_1, \bar{\theta}_2, \bar{\pi}$ , etc. pour représenter les estimations calculées, si  $H_0$  est vraie.

où

$$n^{1/2}(\bar{\theta}_2 - \bar{\theta}_2) \rightarrow N[\bar{0}, (\bar{X}_2'\bar{P}\bar{X}_2)^{-1}]$$

$$X_2' = (I - \bar{X}(\bar{X}'\bar{P}\bar{X})^{-1}\bar{X}'\bar{P})^{-1}\bar{X}_2'$$

3.3 Approximations

Une approximation de la distribution de  $\Sigma \delta_i Z_i^2$  est

$$\Sigma \delta_i Z_i^2 \approx \frac{\Sigma \delta_i}{\chi^2_{(r-1)(c-1)}}$$

comme dans l'équation (2.2). Étant donné que cette statistique est asymptotiquement équivalente à l'expression (3.1), si  $H_0$  est vraie, on peut calculer la moyenne de (3.1) et obtenir

$$\Sigma \delta_i = \sum_{i=1}^r \sum_{j=1}^c d_{ij} (1 - \pi_{i+} \pi_{+j}) - \sum_{j=1}^c d_{(0)j} (1 - \pi_{+j}),$$

où  $d_{ij}$  est l'effet du plan de sondage dans chaque case;  $d_{(0)j}$  et  $d_{(0)}^j$  sont l'effet du plan de sondage sur les fréquences marginales dans chaque rangée et chaque colonne respectivement. Ce sont Rao et Scott (1983) qui ont obtenu cette expression particulièrement simple. Enfin, Fellegi (1980) a proposé une autre approximation:

$$\left( \sum_{i=1}^r \sum_{j=1}^c d_{ij} / rc \right) \chi^2_{(r-1)(c-1)}$$

Exemple 2

Le tableau 2 comprend un tableau de dimension  $4 \times 2$  qui est basé sur des données de l'Enquête Santé Canada. Ce tableau présente un classement recoupé de l'utilisation de médicaments en fonction de la consommation (quatre catégories: 0, 1, 2, 3 + classes de médicaments une période de deux jours) et selon le sexe (hommes, femmes). Ici,  $n = 31,668$ . La valeur brute de  $W_1$  est 774. La correction de Rao et Scott (1981) réduit cette valeur à 437, tandis que celle de Fellegi (1980) la fait chuter à 327. La statistique de Wald,  $W_6$ , est égale à 538. Hidiroglou et Rao (1981) ont constaté que l'approximation de Rao et Scott (1981) produit d'assez bons résultats en comparaison de ceux de l'approximation de Satterthwaite (1946) qui exige une connaissance complète de la matrice de variances-covariances.

4. MODÈLES LOG-LINÉAIRES

4.1 Échantillonnage multinomial

Nous faisons maintenant une extension des résultats de la section précédente à des classements recoupés plus généraux qui suivent une loi multinomiale. Les formules classiques pour ce genre de modèle figurent dans les ouvrages de Bishop, Fienberg et Holland (1975) et Fienberg (1980).

Tableau 2

Classes de médicaments consommés selon le sexe, Canada (1978-79)

Sexe	0	1	2	3 ±	Total
Hommes	0.293	0.134	0.048	0.021	0.496
	Effet du plan	1.56	3.37	1.15	1.38
	Proportion	0.228	0.159	0.072	0.045
	Effet du plan	3.59	3.13	2.85	1.96
Femmes	0.521	0.293	0.120	0.066	1.000
	Effet du plan	6.03	6.46	1.65	2.57
	Proportion	0.521	0.293	0.120	0.066
	Effet du plan	6.03	6.46	1.65	2.57
Total					

\* A cause de la stratification a posteriori selon l'âge et le sexe, ces effets du plan sont nuls.

### 3.2 Autres méthodes d'échantillonnage

Au lieu de supposer que  $n$  suit une loi multinomiale, on suppose maintenant que  $n \mid (p - \pi) \rightarrow \mathcal{N}(0, V)$  où  $p$  est une estimation calculée à partir de données d'enquête et peut dépendre des poids d'échantillonnage et d'autres facteurs de correction. Dans ce cas, Shuster et Downing (1976) et Fellegi (1980) proposent une statistique de Wald construite à partir de  $\{h_p = p - p, p, \dots\}$ . Soit  $\hat{L}_p$  la matrice de dimension  $(a - 1) \times a$  qui a la forme suivante

$$\hat{L}_p = \begin{bmatrix} \hat{L} \\ 0 \end{bmatrix} \quad (3.2)$$

et définissons  $\tilde{H} = (\tilde{L}_p - \pi_R \tilde{L}_T) \otimes (\tilde{L}_c - \pi_C \tilde{L}_T) - (\pi_R \tilde{L}_T \otimes \pi_C \tilde{L}_T)$

On calcule ensuite la statistique de Wald qui s'écrit

$$W_6 = \tilde{h}^T \tilde{H} \tilde{V} \tilde{H}^T)^{-1} \tilde{h},$$

qui, si  $H_0$  est vraie, suit asymptotiquement une loi de type  $\chi^2_{(r-1)(c-1)}$ . Nous pouvons aussi construire une statistique de Wald dans laquelle on définit  $\{f_p = \log p - \log p, -\log p, \dots\}$ . C'est là un cas spécial du modèle log-linéaire qui est présenté à la section 4. On définit deux matrices qui sont respectivement de dimension  $(r - 1) \times r$  et  $(c - 1) \times c$ :

$$\tilde{E}_R = \begin{bmatrix} \tilde{D}_R & 0 \\ 0 & 0 \end{bmatrix}, \tilde{E}_C = \begin{bmatrix} \tilde{D}_C & 0 \\ 0 & 0 \end{bmatrix}.$$

Soit  $\tilde{F} = (\tilde{E}_R - \tilde{E}_R \tilde{L}_T) \otimes (\tilde{E}_C - \tilde{E}_C \tilde{L}_T) - (\tilde{E}_R \tilde{L}_T \otimes \tilde{E}_C \tilde{L}_T)$ . La statistique de Wald qui découle de ces définitions est

$$W_7 = \tilde{f}^T \tilde{F} \tilde{V} \tilde{F}^T)^{-1} \tilde{f}.$$

Comme il a été fait à la section 2 pour le problème de la qualité de l'ajustement, Rao et Scott (1981) ont montré que les distributions sous  $H_0$  des variables  $W_4, LR_2$  et  $W_5$ , sont toutes asymptotiquement équivalentes à la distribution sous  $H_0$  de (3.1). Par conséquent, cette distribution sous  $H_0$  est identique à celle de  $\sum_{i=1}^r \sum_{j=1}^c Z_{ij}^2$ , où  $\{Z_{ij}\}$  est un ensemble de variables indépendantes qui suivent une loi de type  $\chi^2_1$  et les  $\delta_{ij}$  les valeurs propres de

$$(\tilde{P}_1^R \otimes \tilde{P}_1^C)(\tilde{H} \tilde{V} \tilde{H}^T).$$

Cowan et Binder (1978) ont étudié les propriétés des valeurs propres issues d'un échantillon à deux degrés prélevé selon un plan autopondéré pour un tableau de dimension de  $2 \times 2$ . Ces auteurs ont constaté que la valeur propre augmente quand le degré d'indépendance des proportions dans chaque case diminue à l'intérieur des unités primaires d'échantillonnage.

Nous voulons vérifier l'hypothèse d'indépendance

$$H_0: \pi_{ij} - \pi_{i+} \pi_{+j} = 0 \text{ for } 1 \leq i \leq r-1; 1 \leq j \leq c-1,$$

contre

$$H_1: \pi_{ij} - \pi_{i+} \pi_{+j} \neq 0 \text{ pour un couple quelconque } (i, j).$$

Si nous construisons la variable  $h_{ij} = p_{ij} - p_{i+} p_{+j}$  pour  $1 \leq i \leq r-1$  et  $1 \leq j \leq c-1$ , on peut, par un échantillonnage multinomial sous  $H_0$ , exprimer la matrice de variances-covariances asymptotique de  $\tilde{h} = (h_{11}, \dots, h_{1,c-1}, \dots, h_{r-1,1}, \dots, h_{r-1,c-1})^T$  sous la forme  $\tilde{L}_P \otimes \tilde{P}_C$ , où  $\otimes$  est le produit direct de deux matrices. La statistique de Wald devient donc, si  $H_0$  est vraie,

$$W_4 = \tilde{h}^T \tilde{D}^{-1} \tilde{h} \otimes \tilde{P}_R^{-1} \tilde{h}$$

$$= \sum_{j=1}^c \sum_{i=1}^{r-1} (d_{ij} - d_{i+} d_{+j})^2 / (d_{i+} d_{+j} d_{-ij}),$$

qui correspond au test habituel du khi-carré à  $(r-1)(c-1)$  degrés de liberté.

Un autre test, qui est asymptotiquement équivalent à  $W_4$  si  $H_0$  est vraie, est le test du rapport de vraisemblance exprimé par

$$LR_2 = 2n \left[ \sum_{j=1}^c \sum_{i=1}^{r-1} p_{ij} \log p_{ij} - \sum_{j=1}^c p_{+j} \log p_{+j} - \sum_{i=1}^{r-1} p_{i+} \log p_{i+} \right].$$

On peut aussi résoudre ce problème en prenant un cas spécial des procédés décrits par Grizzle, Starmer et Koch (1969). On calcule la statistique du test de Wald avec

$$\{f_{ij} = \log p_{ij} - \log p_{i+} - \log p_{+j}, \text{ for } 1 \leq i \leq r-1, \text{ and } 1 \leq j \leq c-1\}.$$

La matrice de variances-covariances asymptotique de  $\tilde{f} = (f_{11}, \dots, f_{1,c-1}, \dots, f_{r-1,1}, \dots, f_{r-1,c-1})^T$  est  $(\tilde{D}_R^{-1} - \tilde{1} \tilde{1}^T) \otimes (\tilde{D}_C^{-1} - \tilde{1} \tilde{1}^T)$ . On obtient ainsi une variable de Wald qui a la forme suivante:

$$W_5 = \tilde{f}^T \left[ (\tilde{D}_R \otimes \tilde{D}_C) + \frac{\tilde{D}_R \tilde{D}_C}{\tilde{D}_R \tilde{D}_C} \right] \tilde{f}$$

On note que, si  $H_0$  est vraie,  $f_{ij}$  est asymptotiquement équivalent à

$$\frac{p_{ij}}{p_{+j}} - \frac{p_{i+}}{p_{+j}} + 1,$$

de sorte que  $\sum_{j=1}^c \pi_{i+} f_{ij} \doteq \sum_{j=1}^c \pi_{+j} f_{ij} = 0$ . Cette approximation permet de remplacer  $W_5$  par

$$W'_5 = \sum_{j=1}^c \sum_{i=1}^{r-1} d_{ij} f_{ij}^2.$$

Il convient de noter que si  $H_0$  est vraie, les fonctions  $W_4, LR_2$  et  $W'_5$  sont toutes asymptotiquement équivalentes à

$$(3.1) \quad \sum_{i=1}^r \sum_{j=1}^c \frac{\pi_{i+} \pi_{+j}}{(p_{i+} - \pi_{i+} \pi_{+j})^2} - \sum_{i=1}^r \sum_{j=1}^c \frac{\pi_{i+}}{(p_{i+} - \pi_{i+})^2} - \sum_{j=1}^c \sum_{i=1}^r \frac{\pi_{+j}}{(p_{+j} - \pi_{+j})^2}$$

L'utilité de ce résultat sera montrée à la section (3.3).



Tableau 1

Répartition par âge des buveurs de 1 à 6 consommations par semaine.  
Répartition par âge de la population canadienne projetée par le recensement (1978-1979)

Âge	Répartition de la population		Répartition des buveurs de 1 à 6 cons./semaine		Effet du plan de sondage	
	15-19	20-24	25-34	35-44	45-54	55-64
Total	1.000	.127	.218	.152	.140	.115
65 +	.133	.127	.218	.152	.140	.115
15-19	.117	.150	.264	.175	.148	.093
20-24	.14	.12	2.2	1.1	0.6	1.1
25-34						
35-44						
45-54						
55-64						
65 +						

### Exemple 1

À partir des données de l'Enquête Santé Canada (1978-1979), qui est fondée sur un échantillon stratifié à plusieurs degrés, on a calculé la répartition par âge des buveurs de 1 à 6 consommations par semaine dans un échantillon de 5,204 personnes âgées de 15 ans et plus. Une description de cette enquête est présentée dans la publication "La santé des Canadiens" (n° 82-538F au catalogue de Statistique Canada).

Les résultats, que sont extraits d'une étude de Hidiroglou et Rao (1981), figurent au tableau 1. La valeur brute de  $W_1$  est 298 et elle diminue à 248 après le calcul de l'approximation (2.2). Pour ces données, les rectifications par la stratification à posteriori des groupe d'âge et sexe produisent des effets de plan de sondage assez petits.

Une deuxième approximation de la distribution de  $\sum \delta_j Z_j'$  qui a été proposée par Rao et Scott (1981) est l'approximation de Satterthwaite (1946):  $\sum \delta_j Z_j' \approx a \chi_p^2$ . Pour calculer  $a$  et  $p$ , il faut résoudre

$$\sum \delta_j^2 = \text{tr} \{ (P_1^{-1} V)^{-2} \}$$

$$= \frac{\sum_{j=1}^t \sum_{k=1}^K v_{jk}^2}{\sum_{j=1}^t (\sum_{k=1}^K \pi_{jk}^2)}$$

Cependant, cette solution dépend de tous les éléments de la matrice  $V$ . Le fait important ici à souligner est qu'il est nécessaire de corriger la statistique du test multinomial pour établir le seuil critique approprié.

Une autre approximation possible est celle proposée par Fellegi (1980) et qu'on calcule en divisant la statistique  $n(P - \bar{\pi})' \bar{P}_0^{-1} (P - \bar{\pi})$  par la moyenne des effets du plan de sondage,  $\bar{d}$ , au lieu de la moyenne pondérée qui figure dans l'équation (2.2). L'effet de cette modification sur les données du tableau 1 est que la valeur corrigée de la statistique du chi-carré devient 243, ce qui est comparable au résultat qu'on obtient avec l'approximation de Rao et Scott (1981).

## 3. TESTS D'INDEPENDANCE DANS UN TABLEAU A DEUX DIMENSIONS

### 3.1 Echantillonnage multinomial

Nous supposons maintenant que les catégories de la distribution multinomiale peuvent être recoupées pour former un tableau  $r \times c$  dans lequel, pour l'observation du couple de variables  $(Y_1, Y_2)$ , nous avons  $\Pr(X_1 = i, Y_2 = j) = \pi_{ij}$ ;  $\sum_{j=1}^c \pi_{ij} = \pi_{i\cdot}$ ;  $\sum_{i=1}^r \pi_{ij} = \pi_{\cdot j}$ ;  $\sum_{i=1}^r \sum_{j=1}^c \pi_{ij} = 1$ . On définit  $\pi_{i\cdot} = \sum_{j=1}^c \pi_{ij}$  et  $\pi_{\cdot j} = \sum_{i=1}^r \pi_{ij}$  et  $\bar{\pi} = (\pi_{1\cdot}, \dots, \pi_{r\cdot})'$ ,  $\bar{\pi}_C = (\pi_{1\cdot}, \dots, \pi_{r\cdot})'$ ,  $\bar{\pi}_R = (\pi_{\cdot 1}, \dots, \pi_{\cdot c})'$ ,  $\bar{P}_C = \bar{D}_C - \bar{\pi}_C \bar{\pi}_C'$ ,  $\bar{P}_R = \bar{D}_R - \bar{\pi}_R \bar{\pi}_R'$ ,  $\bar{P} = \bar{D} - \bar{\pi}_C \bar{\pi}_C' - \bar{\pi}_R \bar{\pi}_R'$ . Nous observons le vecteur aléatoire  $n$  de la distribution multinomiale, où  $E\{n\} = n\pi$ . On définit  $\bar{p} = \bar{n}/n$ ,  $\bar{p}_{\cdot} = \sum_{j=1}^c \bar{p}_{ij}$  et  $\bar{p}_{\cdot j} = \sum_{i=1}^r \bar{p}_{ij}$ .

Cette approximation est le résultat du fait que, si  $H_0$  est vraie,

$$\pi_{k0}(f_k - \mu_{k0}) = p_k - \pi_{k0}$$

$$= - (\bar{d} - \bar{\pi}_0)^T \bar{1} - (\bar{\mu} - \bar{\mu}_0)^T \bar{\pi}_0.$$

Notons que  $W_2$  est aussi asymptotiquement équivalent à la statistique de test du Khi-carré de Pearson sous l'hypothèse nulle.

## 2.2 Autres méthodes d'échantillonnage

Ces résultats au sujet de  $W_1$ ,  $W_2$  et  $LR_1$  sont bien connus. Le problème qui nous intéresse ici est celui de l'hypothèse plus générale selon laquelle  $n(\bar{d} - \bar{\pi}) \rightarrow N(\bar{0}, \bar{V})$ , où  $\bar{V}$  n'est pas nécessairement égale à  $\bar{P}$ . Ici,  $\bar{d}$  est une estimation de  $\bar{\pi}$  qui est calculée à partir des données d'une enquête et qui peut dépendre des poids d'échantillonnage et d'autres facteurs de correction, ce qui est souvent le cas dans les enquêtes dont le plan de sondage est complexe. Nous supposons que  $\bar{V}$  est un estimateur convergent de  $\bar{V}$ . Nous examinerons maintenant deux problèmes. Premièrement, on peut construire la statistique de Wald appropriée pour le plan de sondage en question. On obtient ainsi

$$W_3 = n(\bar{d} - \bar{\pi}_0)^T \bar{V}^{-1} (\bar{d} - \bar{\pi}_0),$$

où le rang de  $\bar{V}$  est  $k-1$ , de sorte que  $W_3$  suit asymptotiquement une loi de type  $\chi^2_{k-1}$  si  $H_0$  est vraie.

Une deuxième possibilité consisterait à utiliser les fonctions  $W_1$ ,  $W_2$  ou  $LR_1$  directement pour vérifier l'hypothèse nulle. On sait que, selon la théorie de la loi normale à plusieurs variables, la distribution de  $n(\bar{d} - \bar{\pi}_0)^T \bar{P}^{-1} (\bar{d} - \bar{\pi}_0)$  est celle de  $\sum \delta_i' Z_i'$  où  $\{Z_i'\}$  est un ensemble de variables aléatoires indépendantes de type  $\chi^2_1$  et  $\bar{0} = (\delta_1, \dots, \delta_k)^T$  sont les valeurs propres de  $\bar{P}^{-1} \bar{V}$ ; voir Johnson et Kozi (1970, pg. 150). Ce résultat a été démontré par Rao et Scott (1981), qui ont appelé les  $\delta_i$  les effets généraux du plan de sondage. Notons que, pour  $k=2$ ,  $\bar{0} = n\sigma^2/\{\pi_0(1 - \pi_0)\}$ , où  $\sigma^2 = V(p) = V\{p\}$ , ce qui correspond à l'effet habituel du plan de sondage pour  $p$  si  $H_0$  est vraie.

## 2.3 Approximations

En général, la fonction de répartition de combinaisons linéaires de variables aléatoires de type  $\chi^2$  est plutôt compliquée, quoique leurs moments soient faciles à calculer. Rao et Scott (1981) ont proposé deux approximations pour le calcul des seuils des tests. Dans la première approximation, la distribution en cause est considérée comme étant proportionnelle à celle d'une variable aléatoire de type  $\chi^2_{k-1}$ , et on obtient la constante de proportionnalité en égalant la moyenne de la distribution approximative et celle de la distribution théorique. On parvient ainsi au résultat suivant:

$$\sum_{i=1}^{k-1} \delta_i' Z_i' = \left\{ \sum_{i=1}^{k-1} \delta_i' (k-1) \right\} \chi^2_{k-1} \quad (2.2)$$

Or,

$$\sum \delta_i' = \text{tr}(\bar{D}^{-1} \bar{V})$$

$$= \sum_{i=1}^k V_i / \pi_{i0}$$

$$= \sum_{i=1}^k d_i' (1 - \pi_{i0}),$$

La solution dépend donc seulement des effets du plan de sondage dans chaque case,  $\{d_i'\}$ , où  $V_i$  est le  $i^{\text{ème}}$  élément de la diagonale de  $\bar{V}$  et  $d_i' = V_i / [\pi_{i0}(1 - \pi_{i0})]$ . Cette approximation est particulièrement commode lorsque l'on ne dispose pas de la matrice de variances-covariances au complet, mais que les effets du plan de sondage dans chaque case sont connus, ce qui est souvent le cas dans les publications officielles.

On examine ici les problèmes relatifs à l'ajustement de modèles et aux tests d'hypothèses à partir de données qualitatives d'enquêtes complexes. Quand des données sont recueillies en fonction d'un plan de sondage complexe, il est nécessaire de modifier les méthodes classiques décrites par Imrey, Koch et Stokes pour faire des inférences valables. Si les tableaux publiés indiquent l'effet du plan de sondage sur les fréquences des cases et les fréquences marginales, il est possible de produire une approximation de la loi sous l'hypothèse nulle des statistiques des tests proposés sans disposer du fichier. Toutefois, si on dispose du fichier des données, on peut utiliser d'autres procédés qui seront décrits plus bas. Pour donner une idée générale des modifications qu'il faut apporter aux méthodes habituelles, la section 2 aborde le problème classique de la qualité du test de la validité de l'ajustement. Ensuite, la section 3 examine les tests d'indépendance dans un tableau de contingence à deux dimensions. A la section 4, les modèles log-linéaires seront examinés d'une façon générale. Les modèles de régression logistique sont décrits à la section 5. La section 6 résume les travaux en cours pour l'élaboration d'un projetel à Statistique Canada pour les méthodes étudiées ici et la section 7 porte sur la qualité d'application relative à l'Enquête sur la Population Active du Canada est présentée à la section 5.

## 2. QUALITÉ DE L'AJUSTEMENT

### 2.1 Échantillonnage multinomial

Supposons que nous choisissons  $n$  observations indépendantes,  $Y_1, \dots, Y_m$ , qui ont une même distribution discrète à  $k$  catégories, où  $\Pr(Y = i) = \pi_i$ ;  $\sum_{i=1}^k \pi_i = 1$ . Nous observons le vecteur aléatoire  $n = (n_1, \dots, n_k)^t$ , qui suit une loi multinomiale. Notre estimateur de  $\pi = (\pi_1, \dots, \pi_k)^t$  est  $\hat{p} = n/n$ . Cet estimateur est sans biais et sa matrice de variances-covariances est  $\{(\hat{p} - \pi)\pi^{-1}\} = (\hat{p} - \pi)(\hat{p} - \pi)^t/n$ , où  $\hat{p} = \text{diag}\{\pi_1, \dots, \pi_k\}$ . Notons que  $\hat{p}^{-1} = \hat{p}^{-1} + (\hat{1}\hat{1}^t/\pi_k)$ . Asymptotiquement,  $n(\hat{p} - \pi) \rightarrow N(0, \hat{P})$ . Pour une valeur donnée de  $\bar{\pi}_0$ , le problème de la qualité de l'ajustement consiste à tester l'hypothèse:

$$H_0: \bar{\pi} = \bar{\pi}_0,$$

contre l'hypothèse

$$H_1: \bar{\pi} \neq \bar{\pi}_0.$$

Soit  $\tilde{P}_0$  la valeur de  $\tilde{P}$  associée à  $\bar{\pi}_0$ , le statistique de Wald pour ce test est

$$W_1 = n(\bar{d} - \bar{\pi}_0)^t \tilde{P}_0^{-1} (\bar{d} - \bar{\pi}_0) = u^t \left\{ \sum_{i=1}^k (d_i - \pi_{i0})^2 / \pi_{i0} \right\},$$

ce qui n'est pas autre chose que le test du Khi-carré de Pearson. Si  $H_0$  est vrai, cette statistique suit asymptotiquement une loi de type  $\chi_k^2$ . Le test du rapport de vraisemblance pour ce problème repose sur l'expression

$$LR_1 = 2n \sum_{i=1}^k p_i \log(p_i / \pi_{i0}).$$

Etant donné que  $2p_i \log(p_i / \pi_{i0})$  est asymptotiquement équivalent à  $2(p_i - \pi_{i0})^2 / \pi_{i0}$  si  $H_0$  est vraie, on peut voir que le test du rapport de vraisemblance est asymptotiquement équivalent au test du khi-carré de Pearson sous l'hypothèse  $H_0$ .

Un autre test possible de l'hypothèse nulle est obtenu par le vecteur des logarithmes  $\bar{\mu}_0 = \log \bar{\pi}_0$ , et  $\bar{\mu} = \log \bar{p}$ . Si  $H_0$  est vraie,  $\bar{\mu} - \bar{\mu}_0$  est asymptotiquement équivalent à  $D_1(p - \bar{\pi}_0)$ . Par conséquent,  $n(\bar{\mu} - \bar{\mu}_0) \rightarrow N(0, \hat{D}_1)$ , où  $\hat{D}_1 = \hat{1}\hat{1}^t - \hat{1}\hat{1}^t$  sous  $H_0$  et la statistique de Wald est

$$W_2 = (\bar{\mu} - \bar{\mu}_0)^t [\hat{D}_1^{-1} (\bar{\mu} - \bar{\mu}_0)] = (\bar{\mu} - \bar{\mu}_0)^t \bar{\pi}_0^o / \pi_{k0}^o + \sum_{i=1}^{k-1} \pi_{i0}^o (\bar{\mu}_i - \bar{\mu}_{i0})^2,$$

où  $\mu_{k0} = \log \pi_{k0}$  et  $\bar{\mu}_k = \log p_k$ .

## Analyse de données qualitatives d'enquêtes Complexes: Quelques expériences canadiennes<sup>1</sup>

D.A. Binder, M. Gratton, M.A. Hidiroglou,  
S. Kumar et J.N.K. Rao<sup>2</sup>

### RÉSUMÉ

Les tests de la qualité de l'ajustement, les tests d'indépendance dans un tableau de contingence à deux dimensions, les modèles log-linéaires et les modèles de régression logistique sont examinés par rapport aux échantillons d'enquêtes basées sur un plan de sondage complexe. On relève quelques approximations et de l'hypothèse nulle et on présente des exemples à partir de données de l'Enquête Santé Canada et de l'Enquête sur la Population Active du Canada. Il est aussi brièvement question des possibilités de mise en oeuvre d'un progiciel pour les méthodes étudiées.

**KEYWORDS:**  $\chi^2$  statistic; Wald Statistics; Goodness of fit; Independence in two-way tables; Log-linear and logistic regression model.

### 1. INTRODUCTION

Un résumé du développement des techniques modernes d'analyse des données qualitatives a été présenté dans un excellent document d'étude par Imrey, Koch et Stokes (1981). Comme ces techniques ont été conçues pour des échantillons aléatoires d'unités tirées indépendamment de la même fonction de répartition, elles ne s'appliquent pas directement aux échantillons d'enquêtes basées sur un plan de sondage complexe. Koch *et coll.* (1975) et Shuster et Downing (1976) ont élaboré des méthodes asymptotiques qui reposent sur la variable de Wald et permettent de tenir compte du plan de sondage, mais elles exigent qu'on dispose du fichier complet de données ou, au moins, de la matrice des covariances estimées des estimations pour chaque case. Cohen (1976) et Altham (1976) ont proposé un modèle simple de classification automatique et démontré que la variable généralisée de Wald utilisée pour vérifier la qualité de l'ajustement est un multiple de la variable  $\chi^2$  lorsque ce modèle est valide. Brier (1978) a analysé un modèle semblable, mais s'est penché sur des hypothèses générales au sujet des probabilités dans chaque case et il a prouvé qu'un multiple de la variable correspondante de Pearson est asymptotiquement distribué comme une variable aléatoire  $\chi^2$  lorsque le modèle est valide. Fellegi (1980) a réduit la variable  $\chi^2$  à l'aide d'un facteur de correction calculé à partir de la moyenne des estimations des effets du plan de sondage. Fay (1985) a construit des variables  $\chi^2$  et  $G^2$  pour une analyse fondée sur la méthode du jackknife; ces statistiques permettent également de tenir compte du plan de sondage, mais nous oblige de connaître les estimations obtenues dans chaque case au niveau des unités primaires d'échantillonnage. Rao et Scott (1981) ont décrit un moyen de recififier la variable  $\chi^2$  ou ( $G^2$ ) en utilisant l'approximation de Satterthwaite pour la distribution asymptotique de  $\chi^2$  et la matrice complète des covariances estimées.

<sup>1</sup> Cet article est une version augmentée et révisée de celui présenté lors du Séminaire sur les développements récents dans l'analyse des grands fichiers de données tenu à Luxembourg du 16 au 18 novembre 1983. Le séminaire était patronné par l'Office Statistique des Communautés Européennes.

<sup>2</sup> D.A. Binder, Division, des méthodes d'enquête-institutions et agriculture, M. Gratton, Planification et support informatique, M.A. Hidiroglou, Division des méthodes d'enquêtes-entreprises, S. Kumar, Division des méthodes de recensement et d'enquête-ménages, Statistique Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, et J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada.



- HARTLEY, H.O. et RAO, J.N.K. (1962). Sampling with Unequal Probabilities without Replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- HOFMANN, H., DREW, D., CATLIN, G. et MAYDA, F. (1984). A Proposal for a Telephone Survey Development Program. Document interne, Statistique Canada.
- HUANG, E. et ERNST, L. (1981). Comparison of an Alternate Estimator to the Current Composite Estimator in CPS. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 303-308.
- JUDKIN, D.R. et SINGH, R.P. (1981). Using Clustering Algorithms to Stratify Primary Sampling Units. *Proceedings of the Section on Survey Research Methods, American Statistical Association Meetings*, 274-284.
- KEYFITZ, N. (1951). Sampling with Probabilities Proportional to Size: Adjustment for Changes in the Probabilities. *Journal of the American Statistical Association*, 46, 105-109.
- KEYFITZ, N. (1957). Estimates of Sampling Variance where two Units are Selected from each Stratum. *Journal of the American Statistical Association*, 52, 503-510.
- KOSTANICH, D., JUDKIN, D., SINGH, R. et SCHANTZ, M. (1981). Modification of Friedman-Rubins' Clustering Algorithm for Use in Stratified PPS Sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association Meetings*.
- KUMAR, S. (1982). Investigation of the Labour Force Survey Rounding and Release Criteria. Document technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- KUMAR, S. et LEE, H. (1983). Evaluation de l'application d'estimateurs composites à l'enquête sur la population active au Canada. *Techniques d'enquête*, 9, 196-221.
- LEMAITRE, G. (1983). Some Results from the Time and Cost Study. Document technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- MACREDIE, I. (1983). Family Oriented Measures of Employment and Unemployment. Exposé présenté au comité de travail de l'OCCDE sur les statistiques de l'emploi et du chômage.
- MAYDA, F., DREW, D. et LINDEYER, J. (1984). Phase-in of the Redesigned Labour Force Survey Sample. Document technique, (en voie de rédaction), Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- PAUL, E.C. et LAWES, M. (1982). Caractéristiques des ménages répondants et non répondants dans l'enquête sur la population active du Canada. *Techniques d'enquête*, 8, 53-93.
- PLATEK, R. et SINGH, M.P. (1976). Méthodologie de l'enquête sur la population active du Canada. N° 71-526 au catalogue, Statistique Canada.
- PLATEK, R. et SINGH, M.P. (1977). A Strategy for Updating Continuous Surveys. *Metrika*, 25, 1-7.
- RAO, J.N.K. (1975). Unbiased Variance Estimation for Multi Stage Designs. *Sankhya*, série C, 37, 133-139.
- RAO, J.N.K., HARTLEY, H.O. et COCHRAN, W.G. (1962). On a Simple Procedure of Unequal Probability Sampling without Replacement. *Journal of the Royal Statistical Society*, série B, 24, 482-490.
- SINGH, M.P. et DREW, J.D. (1977). Sample Expansion in Self Representing Units of the Canadian Labour Force Survey. Document technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- SINGH, M.P. et DREW, J.D. (1981a). Research Plans for the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association Meetings*.
- SINGH, M.P. et DREW, J.D. (1981b). Redesigning Continuous Surveys in a Changing Environment. *Techniques d'enquête*, 7, 44-73.
- VERMA, R.B.P., BASAVARAJAPPA, K.G. et BENDER, R.K. (1983). The Regression Estimates of Population for Sub-Provincial Areas in Canada. *Techniques d'enquête*, 9, 242-266.



Une fois que le remaniement de l'échantillon sera terminé, une des principales questions examinées dans les recherches futures sur la méthodologie de l'EPA sera l'élaboration d'un plan de sondage double dans lequel une partie de l'échantillon sera convertie en une base de sondage téléphonique à l'aide des techniques de composition de numéros au hasard (CNH). Dans le cadre d'un nouveau programme d'enquêtes téléphoniques (Hofmann, Drew, Carlin et Mayda 1984), des statisticiens planifient à l'heure actuelle un programme pluriannuel qui comprendra des expériences avec la méthode de CNH pour étudier les incidences d'une augmentation du taux de non-réponses sur l'échantillon choisi pour les interviews téléphoniques, des recherches sur les méthodes d'estimation applicables aux plans de sondage doubles et une évaluation des avantages et des désavantages de la centralisation et de la décentralisation des interviews téléphoniques. Un autre projet d'étude portera sur les moyens de mettre à jour économi- quement l'échantillon aréolaire des régions AR au cours de la période intercensitaire.

## REMERCIEMENTS

Les auteurs remercient les membres du comité chargé du remaniement de l'échantillon de l'EPA pour leur appui et leur conseils pendant le déroulement du programme de remaniement et, plus particulièrement, I.P. Fellegi, D.B. Petrie, G.J. Brackstone, R. Platek, I. Macredie, M. Levine, M. Brochu et F. Mayda. Les auteurs remercient également tous les membres de l'équipe de méthodologie, qui ont participé aux travaux de recherche et de mise en application de ce projet et ont rendu possibles les améliorations décrites plus haut. Les auteurs remercient également l'arbitre pour ses observations utiles.

## BIBLIOGRAPHIE

- CHOUHDARY, G.H., LEE, H. et DREW, J.D. (1984). Cost Variance Optimizations for the Canadian Labour Force Survey. Document technique interne (en voie de rédaction), Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- CHOUHDARY, G.H., LEE, H. et SIDA, R. (1984). Variance Estimation for the Redesignated Labour Force Survey. Document technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- DREW, J.D., CHOUHDARY, G.H. et GRAY, G.B. (1978). Some Methods for Updating Sample Survey Frames and Their Effects on Estimation. *Techniques d'enquête*, 4, 225-263.
- DREW, J.D., SINGH, M.P. et CHOUHDARY, G.H. (1982). Evaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active au Canada. *Techniques d'enquête*, 8, 19-52.
- EARWAKER, S. et BELANGER, Y. (1981). Ratio Estimation at the Subprovincial Level for the Labour Force Survey. Document technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- FELLEGI, I.P., GRAY, G.B. et PLATEK, R. (1967). The New Design of the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 62, 421-453.
- FOY, P., BELANGER, Y., DREW, J.D. et JONCAS, M. (1984). Multivariate Clustering Algorithm for Stratifications and its Application to the Canadian Labour Force Survey. Document technique (en voie de rédaction), Division des méthodes de recensement et d'enquête-ménages, Statistique Canada.
- FRIEDMAN, H.P. et RUBIN, J. (1967). On some Invariant Criteria for Grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.
- GHANGURDE, P.D. (1984). Evaluation of LFS Non-Response Adjustment in Household Size Cells. Document technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- GRAY, G.B. (1973). On Increasing the Sample Size (No. of PSUs). Document technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.

Enfin, le fait que la méthode d'estimation choisie pour le remaniement a été modifiée de façon à ce qu'une distribution de la taille des ménages soit incluse à titre provisoire dans la rectification par la méthode du quotient (des estimations démographiques de la taille des familles seront incluses plus tard comme dimension supplémentaire dans le calcul des estimations par la méthode itérative du quotient pour le dernier degré) augmentera la convergence des statistiques sur le marché du travail relatives aux familles et aux personnes, et améliorera les statistiques sur les dépenses et le revenu des familles.

## 10. SOMMAIRE DES CHANGEMENTS ET DES FUTURS PROJETS DE RECHERCHE

La plupart des travaux de recherche pour le remaniement de l'EPA à partir des résultats du recensement de 1981 sont terminés, la mise en oeuvre du nouveau plan de sondage est en cours et certains aspects de la recherche sur les méthodes d'estimation doivent se poursuivre. Fait intéressant, beaucoup d'études faites pour ce programme ont confirmé la validité des techniques et des méthodes employées jusqu'à présent dans l'EPA comme, par exemple, la sélection de deux UPE par strate dans les régions NAR, l'application de la méthode de Rao, Hartley et Cochran et d'un facteur de densité (4 ou 5 ménages par grappe) aux régions AR et le maintien du plan de renouvellement étalé sur six mois. Toutefois, plusieurs recherches ont également permis d'améliorer le programme de renouvellement et le nouveau plan de sondage, concernant la fiabilité des données (voir section 2), la production d'informations plus précises et facilement accessibles et les progrès technologiques.

Parmi les améliorations du programme de remaniement, mentionnons l'utilisation de données du recensement de 1981 au lieu de données recueillies indépendamment sur le terrain pour la mise à jour de l'échantillon des régions AR, la simplification de la délimitation des degrés du plan de sondage et l'automatisation de la stratification et de la formation des UPE. Il sera également possible d'éviter des dépenses par l'introduction progressive d'une bonne partie de l'échantillon remanié. Ainsi, un groupe de renouvellement choisi selon le nouveau plan de sondage remplacera un groupe de renouvellement de l'ancien échantillon, une fois tous les mois au cours d'une période de 6 mois, alors que, dans le passé, l'ancien échantillon continuait d'être renouvelé pendant 3 ou 4 mois en même temps que le nouvel échantillon croissant (Mayda, Drew et Lindeyer 1984). Il a été estimé que le remaniement actuel coûtera \$1,8 million de moins (en dollars de 1983-1984) que le remaniement précédent.

Les principales améliorations qui ont rendu le plan de sondage de l'EPA plus économique qu'apparaient sont l'extension des interviews téléphoniques à la collecte des données pour la période du deuxième au sixième mois d'inclusion dans l'échantillon à l'intérieur des régions NAR, la révision du plan de sondage des régions NAR pour définir des strates urbaines et rurales explicites, l'élimination d'un degré d'échantillonnage dans les régions rurales, l'élaboration d'un schéma général de stratification pour les régions AR et NAR, l'utilisation de chiffres de population repères dans le calcul des estimations et la mise au point d'une meilleure répartition de l'échantillon. Ces améliorations ont suffi pour accroître la précision des données intraprovinciales sans nuire à celle des données relatives aux provinces, même si la taille globale de l'échantillon a diminué d'environ 6% et 7%. La fiabilité a augmenté de 14% en moyenne pour les coefficients de variation du nombre de chômeurs dans la moitié des régions économiques et des RMR où les estimations étaient moins fiables dans l'ancien plan de sondage, et les changements constatés dans les autres régions étaient faibles ou nuls. Les gains d'efficacité seront encore plus importants pour les estimations intraprovinciales du nombre de personnes occupées. En ce qui a trait au coût de l'enquête, la diminution de la taille de l'échantillon et la réduction des frais entraînée par l'extension des interviews téléphoniques permettront d'économiser environ \$0,7 million par année (en dollars de 1983-1984).

Les améliorations structurelles du plan de sondage découlant de facteurs comme les changements dans la stratification, l'élimination d'un degré d'échantillonnage dans les régions NAR, l'utilisation de chiffres repères infraprovinciaux dans le calcul des estimations et la nouvelle répartition de l'échantillon, ont permis non seulement d'accroître la précision des estimations nationales et provinciales, mais également d'atteindre les objectifs visés pour la production de données infraprovinciales. Il devenait alors possible de réduire la taille globale de l'échantillon de l'EPA afin d'affecter des ressources financières à la collecte occasionnelle de données sur certaines autres variables socio-économiques. La taille de l'échantillon a donc été fixée à 51,500 ménages par mois au lieu de 55,000. Cette baisse globale d'environ 6% et 7% a été réalisée par une diminution uniforme de la taille de l'échantillon dans toutes les provinces sauf l'Ile-du-Prince-Édouard. En outre, les coûts unitaires de la collecte des données diminueront à cause de l'augmentation du nombre d'interviews téléphoniques.

## 9. EFFETS DU REMANIEMENT SUR LES ENQUÊTES LIÉES À L'EPA

Pour la plupart des enquêtes-ménages de Statistique Canada, la base de sondage, le plan d'échantillonnage et les systèmes de collecte et de traitement des données de l'EPA représentent des instruments qui permettent de recueillir des données plus rapidement, plus économiquement et avec plus de précision que par des sondages indépendants. Le plan et les opérations de ces enquêtes sont intégrés, à divers degrés, à ceux de l'EPA.

La plupart de ces projets prennent la forme d'enquêtes supplémentaires pour lesquelles des questions sont ajoutées au questionnaire de l'EPA et les seuls frais sont les coûts additionnels de ces nouvelles questions. Les enquêtes qui peuvent nuire à l'EPA, soit parce qu'elles portent sur un sujet délicat, soit parce que leur questionnaire est trop long, ne sont pas menées de cette manière. Normalement, les données pour ce genre d'enquête sont recueillies par les interviewers de l'EPA, auprès d'un échantillon séparé de ménages, dans les régions où l'EPA est menée. D'autres enquêtes sont encore moins étroitement liées à l'EPA parce qu'elles n'ont pas lieu dans les mêmes régions que l'EPA, mais sont élaborées à partir du plan de sondage de l'EPA et soumises à un contrôle pour éviter un chevauchement avec l'échantillon de l'EPA.

Tel qu'il a été mentionné plus haut, le programme de remaniement vise avant tout l'EPA, mais on a également essayé d'accroître l'utilité générale de l'EPA pour d'autres activités statistiques. Les modifications apportées à l'EPA à cette fin sont résumées brièvement ci-dessous. La redistribution d'une partie de l'échantillon des régions NAR vers les régions AR produira une répartition plus robuste en général et, en particulier, améliorera l'estimation des variations du revenu et des loyers dans l'enquête sur les finances des consommateurs et l'enquête sur les loyers. Par ailleurs, la taille minimale de l'échantillon des RMR produit des estimations relatives aux RMR. Le nouveau schéma général de stratification multidimensionnelle en fonction de 15 variables (dans toutes les régions NAR et AR) sera également mieux pour les autres enquêtes que la méthode actuelle de stratification par activité économique ou par unité géographique.

Trois changements constituent un avantage direct pour les enquêtes menées auprès de différents ensembles de ménages: (i) l'élimination d'un degré d'échantillonnage dans les régions rurales entraînera une réduction énorme du délai de préavis, qui était de 13 mois auparavant, mais de 7 mois dans la nouvelle EPA, (ii) les UPE seront plus compactes géographiquement parce qu'il y aura dorénavant des strates rurales et urbaines explicites, ce qui facilitera les petites enquêtes où les UPE doivent bien correspondre aux tâches des interviewers et (iii) la fiabilité découlant de l'amélioration du programme de stabilisation de l'échantillon permettra de tirer des sous-échantillons de pratiquement n'importe quelle taille, à l'échelle nationale, provinciale ou infraprovinciale, pour les enquêtes élaborées à partir du plan de sondage de l'EPA.



7.6 Nouvelle règle pour l'arrondissement et la publication des chiffres

Dans l'ancienne EPA, les estimations de niveaux étaient arrondies au millième et publiées si elles dépassaient 4,000. Cette règle était appliquée uniformément à toutes les provinces et à toutes les estimations, de manière à ce que les données publiées aient un coefficient de variation de 33,3% ou moins.

Des normes plus rigoureuses pour l'arrondissement et la publication des chiffres de chaque province ont été établies pour l'échantillon remanié. D'ores et déjà, le CV des estimations non arrondies doit être inférieur ou égal à 33,3% et l'erreur d'arrondissement ne doit pas dépasser 20% de l'erreur-type d'une estimation non arrondie. On a constaté que la norme de publication pouvait être réduite à 2,000 ou 3,000 pour toutes les provinces, sauf le Québec et l'Ontario, et que les estimations infraprovinciales devaient être arrondies au centième plutôt qu'au millième (Kumar 1982).

8. REDISTRIBUTION DE L'ÉCHANTILLON

Les objectifs du remaniement visant l'amélioration des données infraprovinciales ont été précisés à la section 2. Les modifications apportées aux techniques et aux méthodes utilisées dans l'EPA favorisent une amélioration générale de la précision des données, cependant, il a également fallu songer à une nouvelle répartition de l'échantillon à l'intérieur des provinces pour atteindre les objectifs (iii), (iii) et (iv). Il était nécessaire d'élargir l'échantillon dans 13 des 66 régions économiques, dans 6 des 24 RMR et dans 27 des 42 villes autres que des RMR. Les CV du nombre de chômeurs se situaient dans l'intervalle de 15% à 25%, la nouvelle répartition a produit une baisse de 12% dans les CV. Pour des raisons d'ordre pratique, dans l'échantillon remanié, les données mensuelles relatives aux RE et aux RMR et les données trimestrielles relatives aux autres villes seront fondées sur des échantillons comprenant un minimum de 300 et 120 ménages respectivement par mois. Il convient de souligner que deux utilisations importantes des données de l'EPA par les ministères fédéraux n'ont pas été prises en considération directement dans les objectifs de remaniement. Il s'agit de la production de moyennes mobiles du taux de chômage de 3 mois consécutifs, dans les régions infraprovinciales de l'assurance-chômage (AC), pour le calcul du nombre de semaines qu'une personne dans chaque région différente doit avoir travaillées pour toucher des prestations d'AC, et de la production de moyennes de 3 ans, dans quelques 180 à 200 régions composées de divisions de recensement distinctes ou combinées, pour la répartition de l'aide financière fédérale aux nouvelles initiatives industrielles. Toutefois, la redistribution de l'échantillon comporte des avantages indirects pour ces deux opérations.

Pour déterminer la taille des échantillons qui convient aux objectifs visés, on a utilisé les taux de chômage moyens de la période 1980-1982, puisque les prévisions à moyen terme indiquent que le chômage demeurera élevé au cours des années 1980.

Une des conséquences générales de la redistribution de l'échantillon a été un déplacement d'une bonne partie de l'échantillon des grandes RMR et des régions économiques vers les petites. Ce transfert s'est répercuté négativement sur les estimations provinciales et nationales parce que l'échantillon n'était plus représentatif proportionnellement. Cette diminution de la précision des estimations globales a cependant été plus que compensée par les gains de la fiabilité dus aux améliorations structurelles que les travaux de recherche ont permis d'apporter aux techniques et aux méthodes de l'enquête.

Une étude a également été menée à l'aide du modèle coût-variance proposé par Fellner, Platt et Gray (1967) pour déterminer les taux de sondage optimaux dans les régions NAR et AR. Le résultat a été un déplacement de l'échantillon des régions NAR vers les régions AR. Ce transfert a été particulièrement important au Québec et en Ontario, où la proportion de l'échantillon dans les régions AR a augmenté de 0,60 à 0,72 (0,78 de la base de sondage), et a permis d'améliorer les estimations provinciales du nombre de chômeurs avec l'équivalent d'une réduction de 5% de la variance pour un échantillon de taille fixe. Cette optimisation a également aidé à atteindre les objectifs (ii) et (iv) et a eu des retombées positives pour l'enquête sur les finances des consommateurs et l'enquête sur les loyers.

régions ont été évaluées. Une des solutions proposées était un estimateur dépendant de l'échantillon, qui combine un estimateur pour domaines stratifiés à posteriori et un estimateur synthétique. Cet estimateur devient exclusivement un estimateur pour domaines stratifiés à posteriori, si la taille de l'échantillon prélevé dans un domaine est suffisante selon certains critères, mais ajoute une composante synthétique dont le poids relatif dépend des lacunes de l'échantillon qui représente un domaine. Selon les résultats d'une étude, il a été recommandé qu'un estimateur dépendant de l'échantillon soit élaboré pour la production d'estimations moyennes annuelles ou pluriannuelles relatives à des régions telles que les CBEF et les DR (Drew, Singh et Choudhry 1982). D'autres travaux d'application, de recherche et de mise au point sont actuellement en cours à Statistique Canada dans le cadre du programme de données sur les petites régions.

#### 7.4 Estimation de la variance

La méthode utilisée pour l'estimation de la variance dans l'échantillon remanée sera de nouveau celle de Keyfitz (1957), quoiqu'elle sera modifiée de façon à produire une estimation à deux étapes par la méthode du quotient pour le degré final, c'est-à-dire une estimation calculée en une seule itération de la méthode itérative du quotient. Comme les itérations subséquentes n'ont qu'un très petit effet sur les estimations, on n'en tient pas compte dans l'estimation de la variance. D'autres modifications de la technique actuelle d'estimation de la variance sont également à l'étude. Ainsi, on songe à utiliser les grappes comme échantillons répétés dans les UPE, au lieu de regrouper les grappes en deux pseudo-échantillons répétés, comme à l'heure actuelle.

On a également songé à remplacer la méthode actuelle d'estimation de la variance par celles décrites par Rao, Hartley et Cochran (1962) et Rao (1975) pour les UPE, où le plan de sondage de Rao, Hartley et Cochran est utilisé (Choudhry, Lee et Sida 1984). La méthode actuelle ainsi que les autres ont été examinées avec ou sans correction par la méthode du quotient. On a constaté que, sans correction par la méthode du quotient, la méthode actuelle exagère la variance de certaines caractéristiques (par exemple, la variance du nombre de personnes occupées est surestimée de 20%), mais, avec l'estimation par le quotient, les biais sont négligeables. Les biais estimés étaient également négligeables dans le cas des autres estimateurs comparés. Le principal avantage des ces autres estimateurs est qu'ils sont plus stables. Toutefois, on a retenu la méthode actuelle parce qu'elle est simple, alors qu'il est complexe d'estimer les variances des variations ou des moyennes avec les autres méthodes.

#### 7.5 Estimateurs composites

Dans l'EPA, il existe des corrélations modérées ou fortes entre la plupart des caractéristiques observées au cours de deux mois consécutifs, parce que 5/6 de l'échantillon demeure le même. Kumar et Lee (1983) ont étudié divers estimateurs composites qui mettent à profit ces corrélations en utilisant des données sur les échantillons précédents pour améliorer les estimations du mois courant. Leurs expériences ont porté sur une classe d'estimateurs composites AK examinées récemment par Huang et Ernst (1981) et sur d'autres estimateurs employés dans la Current Population Survey du U.S. Bureau of the Census.

En supposant que l'estimateur de la méthode du quotient est sans biais, Kumar et Lee ont évalué l'erreur quadratique moyenne et constaté qu'un compromis entre les poids A et K produisait un gain d'efficacité de 50% au plus pour les estimations mensuelles du niveau de l'emploi et du chômage, et de 50% à 160% pour les estimations correspondantes des variations mensuelles. Avant qu'une décision soit prise concernant l'utilisation d'estimateurs composites dans l'EPA, d'autres études doivent être faites sur ces estimateurs pour évaluer l'effet des variations dans le biais de renouvellement dues aux modifications apportées aux rectifications pour tenir compte de la non-réponse et au calcul des estimations par le quotient et on se penchera également sur les répercussions possibles de ces estimateurs sur le déroulement de l'enquête.



## 6.6 Échantillonnage à deux degrés pour l'Île-du-Prince-Édouard

Pour la plus petite province du Canada, l'Île-du-Prince-Édouard, les fractions de sondage ont été fixées à 4% pour que les estimations mensuelles sur la population active aient le degré d'exactitude nécessaire. Comme ces taux de sondage sont élevés, on a adopté un plan de sondage moins complexe fondé sur un échantillon à deux degrés composé de SD et de logements et stratifié en fonction du sexe et de l'âge. On a constaté que ce plan de sondage est un peu plus coûteux que  $P_{10}$ , mais, en général, si on tient également compte de la variance, il est bien supérieur à  $P_{10}$ , l'efficacité relative de l'estimation du nombre de personnes occupées et de chômeurs étant de 2,21 et 1,11 respectivement par rapport à celles de  $P_0$  (Choudhry, Lee et Drew 1984).

## 7. ESTIMATION

### 7.1 Calcul d'estimations par la méthode du quotient pour le degré final

Dans l'ancienne version de l'EPA, des estimations détaillées pour les diverses tranches d'âge et sexe dans chaque province étaient calculées par la méthode du quotient pour le degré final. Depuis que Statistique Canada a élaboré des méthodes permettant de produire des estimations démographiques infraprovinciales de meilleure qualité et dans des délais plus courts (Verma, Basa et Bender 1983), une étude a été entreprise sur une étape intermédiaire d'estimation par le quotient dans laquelle des estimations d'enquête sur la population âgée de 15 ans et plus sont corrigées par la méthode du quotient à partir d'estimations extérieures, avant le calcul des estimations finales par la méthode du quotient. On a constaté que cette étape intermédiaire n'avait pas d'effet sur les variances des données relatives aux provinces, mais permettait de réduire la variance des estimations relatives aux régions infraprovinciales de près de 20% pour le nombre de personnes occupées et de 74% pour le nombre de chômeurs (Eardwaker et Bélanger 1981). En pratique, une technique d'estimation itérative par la méthode du quotient, où deux étapes d'estimation par le quotient sont répétées jusqu'à ce que les deux conditions de contrôle soient satisfaites, a été appliquée au début de 1983.

### 7.2 Amélioration des estimations relatives aux ménages et aux familles

Paul et Lawes (1982) ont utilisé les fichiers de données longitudinales de l'EPA, qui englobent les ménages pour les six mois de leur inclusion dans l'échantillon, pour démontrer que les taux de non-réponse les plus élevés sont observés chez les ménages composés d'un petit nombre de membres. Dans l'ancienne EPA, la non-réponse était compensée par une nouvelle pondération des données sur des régions. Cette correction était faite sans tenir compte de la taille des ménages, de sorte que les estimations de la taille des ménages et des familles renfermaient des biais de 1% à 3%. Il y avait également un autre problème qui entraînait un manque de convergence dans les statistiques sur les familles et les personnes (Macredie 1983). La Division de la démographie de Statistique Canada est actuellement en train d'établir des estimations démographiques de la taille des familles et, une fois que ces résultats seront connus, ils seront intégrés au calcul des estimations par la méthode itérative pour le degré final, afin de résoudre ces deux problèmes. Provisoirement, une étude a été entreprise pour estimer les distributions de la taille des ménages répondants et non répondants à l'aide des données longitudinales de l'EPA, avant le calcul de l'estimation finale par la méthode itérative du quotient (Changurde 1984).

### 7.3 Estimations pour les petites régions

Les demandes d'estimations de la population active dans les petites régions (domaines) telles que les circonscriptions électorales (CE) et les divisions de recensement (DR), qui représentent chacune plus de 250 unités d'un bout à l'autre du Canada, ont augmenté au cours des dernières années. Comme il n'a pas été possible d'utiliser exactement les mêmes limites que celles de ces régions dans le plan de l'enquête, diverses autres techniques d'estimation pour les petites

### 6.3 Plan de sondage avec stratification urbaine et rurale

L'ancien plan de sondage renfermait une stratification urbaine et rurale implicite. Les UPÉ étaient définies de manière à ce que le rapport entre la population urbaine et rurale soit à peu près le même que celui de la strate et, à l'intérieur des UPÉ sélectionnées, les parties urbaines et rurales étaient échantillonnées indépendamment. Théoriquement, chaque UPÉ devait correspondre à la tâche d'un intervieweur. Toutefois, en pratique, cette correspondance était floue parce que, comme il fallait le rapport voulu entre la population urbaine et rurale, il arrivait souvent que les parties urbaines et rurales des UPÉ n'étaient pas contiguës.

L'ancien plan de sondage,  $P_0$  (dans lequel les grappes rurales étaient supprimées), a été comparé avec un autre plan de sondage,  $P_1$ , qui renfermait une stratification urbaine et rurale explicite. Comme  $P_0$ ,  $P_1$  comprenait 3 étapes d'échantillonnage, tant dans les régions urbaines que dans les régions rurales. Les strates urbaines étaient ventilées en UPÉ (centres urbains distincts ou voisins), en grappes et en logements; les strates rurales étaient classées en UPÉ (groupes de SD ruraux voisins), en unités secondaires d'échantillonnage (SD) et en logements. Dans  $P_1$ , les UPÉ urbaines et les UPÉ rurales étaient indépendamment de manière à produire des échantillons correspondant à des tâches d'interviewer.

Ces deux plans ont été évalués en fonction de la variance d'échantillonnage et du coût d'enquête (Choudhry, Lee et Drew 1984). Dans l'étude de la variance, ces deux plans ont été simulés pour des strates comprenant 2 UPÉ à l'aide de chiffres repères du recensement de 1971 et les variables étudiées étaient tirées des données du recensement de 1976.

Pour comparer les coûts de ces plans de sondage, un modèle simple a été élaboré pour l'exploitation de  $P_0$ . L'ancien plan de sondage, par des interviews téléphoniques, et les différents coûts ont été estimés à partir des résultats d'une étude détaillée des temps et des coûts (Lemaitre 1983). Pour estimer le coût relatif des déplacements nécessaires par les plans  $P_0$  et  $P_1$ , on a fait une étude de simulation dans laquelle la dispersion moyenne de l'échantillon correspondant à ces deux plans a été mesurée jusqu'au deuxième degré du plan de sondage à partir des centroïdes de la population des SD. Il s'est avéré que le rendement du plan  $P_1$  était 1,09 fois celui de  $P_0$  et que, sur le plan des coûts d'enquête et de la variance d'échantillonnage,  $P_1$  était supérieur à  $P_0$ . En outre, l'efficacité relative de  $P_1$  par rapport à  $P_0$  était de 1,25 pour le nombre de personnes occupées et de 1,05 pour le nombre de chômeurs.

Compte tenu de ces résultats, le plan  $P_1$  a été appliqué dans 70% des régions économiques où la population urbaine et rurale était suffisante pour former des strates séparées. Dans les autres régions économiques, sauf à l'Île-du-Prince-Édouard, le plan  $P_0$  a été adopté (voir section 6.6).

### 6.4 Nombre d'UPÉ choisies à l'intérieur de chaque strate

Dans le plan de sondage de l'EPA, comme la taille de l'échantillon de chaque UPÉ est fixe, le nombre d'UPÉ choisies à l'intérieur de chaque strate détermine également le nombre de strates. Dans plus de deux tiers des cas, les strates urbaines, rurales ou combinées à l'intérieur des UPÉ ne produisaient que l'équivalent de l'échantillon de 2 ou 3 UPÉ. On a exclu la possibilité de stratifier davantage ces régions parce qu'il devrait y avoir au moins 2 UPÉ par strate pour permettre l'estimation sans biais de la variance.

Les autres RE ont été stratifiées de manière à obtenir 2 ou 3 UPÉ par strate. La réduction estimée de la variance du premier degré par rapport à celle de l'ancien plan de sondage, où il y avait de 3 à 6 UPÉ par strate, atteint 14% pour le nombre de personnes occupées (Choudhry, Lee et Drew 1984). La stratification repose sur l'algorithme de classification décrit à la section 5.

### 6.5 Utilisation de l'algorithme de classification pour la formation des UPÉ des régions NAR

Dans l'ancien et le nouveau plan de sondage de l'EPA, la stratification a lieu avant la formation des unités primaires d'échantillonnage des régions NAR. Les UPÉ sont délimitées de manière à ressembler autant que possible à l'ensemble de leur strate selon les variables de stratification et à être aussi compactes que possible sur le plan géographique. Les UPÉ sont délimitées à l'aide de l'algorithme de classification résumé plus haut et reposent sur une minimisation dans le cas des variables géographiques et une maximisation dans le cas des variables non géographiques.

On a procédé à un premier essai sur le terrain visant uniquement les régions urbaines où plus de 80% des lignes sont privées. Ce test a été appliqué à une partie de l'échantillon réel de l'EPA pour évaluer l'effet des interviews téléphoniques sur la qualité des données. Pour faciliter cette analyse, les tâches des interviewers ont été divisées en une partie qui devait être faite par téléphone et une autre qui devait être faite sur place.

Cette expérience a eu lieu de janvier 1982 à juin 1983 et a été réalisée progressivement de façon à ne pas nuire au déroulement de l'enquête. Les principaux résultats démontrent que les taux de non-réponses sont moins élevés avec les interviews téléphoniques (3,4% contre 4,3% dans l'échantillon témoin), qu'une proportion élevée de ménages ont un téléphone (96% dans toutes les provinces sauf une), que peu de ménages (1%) refusent d'être interviewés par téléphone et que les interviews téléphoniques ne produisent aucune différence notable entre les estimations des caractéristiques de la population active.

Un deuxième essai dans les régions rurales a abouti à des conclusions semblables. Comme les résultats de ces deux essais étaient positifs, il a été décidé de mener des interviews téléphoniques dans l'ensemble des régions NAR pour le reste de l'année 1983 et le début de 1984. Les principales conséquences de cette décision sur le plan de sondage de l'échantillon des régions NAR ont été les suivantes:

- (i) Augmentation du volume des tâches des interviewers: Dans l'ancien plan de sondage, les tâches dans les régions NAR comprenaient en moyenne 50 logements. Comme les coûts unitaires sont faibles dans les tâches qui englobent beaucoup de ménages (Lemaitre 1984) et que l'utilisation du téléphone réduit les déplacements, il a été possible de choisir de 55 à 60 ménages pour l'échantillon de chaque UPE des régions NAR.
- (ii) Attribution du même numéro de renouvellement au deuxième niveau: Dans le nouveau plan de sondage, contrairement à l'ancien plan, le même numéro de renouvellement sera attribué à tous les logements dans les unités secondaires, ce qui permettra de réduire le nombre de visites dans les unités secondaires, du deuxième au sixième mois d'inclusion d'un ménage dans l'échantillon.

## 6.2 Élimination d'un degré d'échantillonnage dans les régions rurales

Dans l'ancien plan de sondage, la sélection de l'échantillon rural à l'intérieur des UPE se faisait en trois étapes: unités secondaires (secteurs de dénombrement du recensement), grappes et logements. Les grappes correspondaient à des zones de terrain faciles à délimiter, qui comprenaient jusqu'à 20 logements et étaient définies à partir de chiffres relevés sur le terrain quand une nouvelle unité secondaire était ajoutée à l'échantillon. À l'intérieur des unités secondaires, on choisissait généralement 5 ou 6 grappes composées de 3 ou 4 logements.

Dès le début du programme de remaniement, on a songé à supprimer la sélection de grappes dans les régions rurales pour les raisons suivantes: (i) la variance d'échantillonnage diminuerait parce qu'il y aurait un degré de moins dans le plan de sondage et (ii) le délai de préavis nécessaire pour inclure de nouvelles unités secondaires dans l'échantillon diminuerait de 13 mois à 7 mois. Une étude a été menée sur le terrain auprès d'un échantillon d'unités secondaires, au moment de leur inclusion dans l'échantillon de l'EPA, pour évaluer la possibilité de tenir à jour de bonnes listes des logements dans les SD ruraux et pour examiner les coûts d'un tel programme. Les résultats obtenus pour ces deux questions ont été positifs. L'effet que l'élimination de cette étape de l'échantillonnage pourrait avoir sur la variance a également été étudié. L'ancien plan de sondage et le plan de sondage modifié ont été simulés à l'aide des données du recensement de 1971 et les composantes de la variance ont été mesurées pour l'estimateur de Horwitz-Thompson sans estimation par le quotient. Le plan de sondage modifié a permis de réduire la variance totale de 20% à 25% pour les principales caractéristiques de la population active (Choudhry, Lee et Drew 1984). Compte tenu de ces résultats, on a décidé, au début du programme de remaniement, d'éliminer l'étape de la sélection des grappes dans les régions rurales et d'examiner d'autres aspects plus généraux du plan de sondage.



Tableau 1  
Pourcentage de réduction de la variance du  
premier degré attribuable à la stratification  
Comparaison de l'ancienne et de la nouvelle méthode

Variable	Méthode de stratification		Variable	Méthode de stratification	
	ancienne	nouvelle		ancienne	nouvelle
total, personnes occupées	9.1	12.1	agriculture <sup>1</sup>	5.9	3.9
revenu d'emploi	18.1	30.4	forêts/pêche <sup>1</sup>	3.1	2.4
personnes ayant terminé leurs études secondaires	39.4	42.1	industries minières <sup>1</sup>	4.8	3.0
population de 15 ans et plus	9.2	12.6	industries manufacturières	23.5	23.1
population de 15 à 24 ans	12.9	17.6	construction	11.9	11.2
population de 55 ans et plus	25.3	29.7	transports	4.2	6.4
total, logements	28.5	33.1	services	14.5	19.8
logements loués	20.9	28.8	nombre de chômeurs <sup>1</sup>	7.1	9.7
ménages d'une seule personne	33.7	38.4			
ménages de deux personnes	27.5	29.6			

<sup>1</sup> caractéristiques non utilisées pour l'optimisation, dans la nouvelle méthode.

Dans les régions NAR, le même algorithme de classification a été appliqué à chaque région économique pour définir des strates rurales ou des strates mixtes urbaines et rurales (voir section 6.3). L'utilisation de l'algorithme de classification pour la formation des UPÉ est examinée à la section 6.5.

## 6. ASPECTS DU PLAN DE SONDAGE DES RÉGIONS NAR

### 6.1 Extension des interviews téléphoniques

Depuis le début des années 1970, la collecte des données pour la période du deuxième au sixième mois d'inclusion d'un ménage dans l'échantillon se fait par interview téléphonique, dans les régions autoréprésentatives, essentiellement pour réduire les coûts. Par contre, dans les régions NAR, les interviews sont encore effectuées par des interviewers à cause des risques que le grand nombre de lignes partagées représente pour le caractère confidentiel des renseignements demandés. Toutefois, comme les interviews téléphoniques permettent de réaliser des économies immédiates et de tirer des avantages à long terme de techniques telles que la composition de numéros au hasard et les interviews téléphoniques assistées par ordinateur, on a décidé d'examiner la possibilité d'étendre les interviews téléphoniques aux régions NAR.

Toutes les variables correspondant à des branches d'activité comprenant moins de 2% de la population active d'une région en train d'être stratifiée ont été supprimées, et le poids des autres variables est modifié pour qu'elles aient une importance égale dans le processus d'optimisation. Le nombre de chômeurs n'a pas été inclus parmi les variables de stratification à cause de son instabilité. Une étude a démontré que, si on les évalue après le recensement subséquent, les strates définies sans tenir compte du nombre de chômeurs sont plus efficaces non seulement pour l'estimation d'autres caractéristiques, mais également pour le calcul du nombre de chômeurs. En revanche, l'inclusion de variables sur les différents quartiers a permis d'accroître l'efficacité des estimations du nombre de chômeurs.

## 5.2 Stratification à deux niveaux dans les régions AR

Dans les grandes UAR dont l'échantillon englobe 300 ménages ou plus, deux niveaux de stratification ont été définis. Les strates primaires, dont l'échantillon aréolaire et l'échantillon d'immeubles d'appartements comprennent de 150 à 170 ménages en tout, sont des ensembles de secteurs de recensement contigus. Les strates primaires sont définies de manière à correspondre à deux tâches d'interviewer. Trois ou quatre strates secondaires aréolaires non géographiques comprenant chacune six ou un multiple de six grappes échantillonnées sont créées dans chaque strate primaire; les secteurs de recensement servent d'unités de stratification et l'optimisation est fondée sur les données du recensement de 1981 relatives aux caractéristiques des personnes n'habitant pas un immeuble d'appartements.

L'échantillon d'appartements est prélevé séparément par échantillonnage systématique avec ppt à partir d'une base de sondage ouverte et comprend généralement une seule strate pour l'ensemble d'une UAR. Un tri des appartements existant au moment de la conception du plan de sondage, en fonction des strates primaires, a produit une stratification géographique implicite pour l'échantillon d'immeubles d'appartements.

Dans les petites UAR décomposables en côtés d'îlots où ni un échantillon séparé d'immeubles d'appartements, ni des strates géographiques primaires n'étaient justifiés, des strates aréolaires non géographiques optimales ont été construites directement. Dans les villes pour lesquelles on ne disposait pas de données agrégées au niveau des côtés d'îlots, la stratification était beaucoup moins facile; on a décidé de former des strates géographiques simples.

Les deux niveaux de stratification définis pour les grandes UAR semblaient appropriés des points de vue opérationnel et technique. Comme les contraintes géographiques sont moins strictes que celles de l'ancien plan de sondage, on a pu accroître l'optimalité des résultats; le fait de consacrer la contrainte de contiguïté aux niveaux élevés permettra d'avoir une unité convenable pour la mise à jour de l'échantillon. Plus tard au cours des années 1980, et facilitera la planification des tâches des interviewers. Par ailleurs, dans l'ancien plan de sondage, les strates AR étaient généralement confiées à un seul interviewer; les estimations de la variance ne tenaient donc pas compte de la composante de la variance totale attribuable à la variance de réponse corrélée. Comme les tâches d'interviewer sont généralement divisées géographiquement et que les strates secondaires sont non géographiques, cela permet d'avoir une superposition des strates et des tâches d'interviewer dans le nouveau plan de sondage, sans augmenter les coûts de la collecte des données, et de mieux prendre en considération la variance de réponse corrélative dans les estimations de la variance. Le tableau 1 présente les résultats d'une étude de variance pour deux UAR—Ottawa et la ville de Québec—et permet de comparer l'efficacité des strates géographiques utilisées dans l'ancien plan de sondage avec celle des strates optimales à deux niveaux pour une expérience faite à partir des données du recensement de 1971. Les chiffres calculés au moment du recensement de 1981 relatifs vement au pourcentage de réduction de la variance du premier degré attribuable à la stratification indiquent que les plus grands changements découlent de la stratification optimale pour le revenu et les logements loués. Le fait que l'amélioration ne soit que légère pour d'autres caractéristiques telles que le nombre de personnes occupées et de chômeurs confirme la force et la robustesse de la stratification simple, mais profonde, de l'ancien plan de sondage.



Villes non décomposables en côtés d'îlots

Comme plus de 70% des UAR non décomposables en côtés d'îlots étaient nouvelles ou n'avaient plus les mêmes limites qu'auparavant et que la plupart des autres UAR n'avaient pas été mises à jour depuis le remaniement de 1973, on a décidé de remanier au complet l'échantillon de ces villes. La grappe définie pour ces UAR était un îlot de recensement ou une combinaison d'îlots dans les zones bâties de ces villes et les chiffres sur les nombres de logements étaient tirés directement des registres des visites et des cartes géographiques utilisés lors du recensement. À l'extérieur des zones bâties, des SD complets ou divisés ont été choisis comme grappes et des comptes relevés sur le terrain ont parfois été utilisés pour diviser des SD.

L'utilisation des registres des visites du recensement a coûté plus que le système employé pour les villes décomposables en côtés d'îlots, mais nettement moins que les comptages indépendants sur le terrain qui étaient nécessaires à l'ancien plan de sondage.

5. STRATIFICATION

5.1 Algorithme et variables de stratification

Une version modifiée d'un algorithme non hiérarchique élaboré par Friedman et Rubin (1967) a été adoptée pour la stratification des régions AR et NAR. Ce choix repose sur les résultats d'études faites par Juddkins et Singh (1981) et Kostianich, Juddkins, Singh et Schanz (1981), qui ont évalué plusieurs algorithmes de stratification pour l'enquête sur la Current Population Survey du U.S. Bureau of the Census. Les modifications apportées à cet algorithme permettent de former des strates géographiquement contiguës et/ou compactes et offre la possibilité de construire soit des grappes homogènes (c'est-à-dire des strates), soit des grappes hétérogènes (c'est-à-dire des unités primaires d'échantillonnage à l'intérieur des strates NAR). Une description détaillée de cette méthode et des résultats d'évaluations empiriques a été présentée par Foy, Belanger, Drew et Joncas (1984). Cette section en donne un résumé.

L'algorithme répartit d'abord les unités au hasard en un nombre déterminé de strates. À chaque itération, toutes les unités de stratification sont examinées et ajoutées à une strate quelconque de manière à minimiser la somme pondérée des carrés de plusieurs variables à l'intérieur de la strate sans dépasser les contraintes relatives à la taille des strates. L'algorithme atteint un optimum local lorsque l'addition d'une unité augmente cette somme des carrés à l'intérieur d'une strate. Une étude de Juddkins et Singh (1981) a révélé que les optimums locaux sont meilleurs quand le nombre d'origines aléatoires est assez élevé (c'est-à-dire 30).

Pour l'option de contiguïté, l'algorithme lit une matrice qui précise, pour chaque unité, toutes les autres qui y sont contiguës. Les premières strates contiguës qui satisfont aux contraintes relatives à leur taille sont constituées à partir des unités choisies au hasard comme points de départ. Pendant l'optimisation, le transfert des unités se fait sous une nouvelle condition selon laquelle la contiguïté doit être maintenue. Pour que les strates soient compactes, les centroïdes de la population (longitude et latitude) sont ajoutés comme variables dans la somme pondérée des carrés qui doit être minimisée.

La stratification des régions NAR et AR repose sur les données du recensement de 1981 et utilise jusqu'à 17 variables de stratification. Parmi les variables démographiques utilisées, il y a le nombre total de personnes occupées, le revenu d'emploi, le nombre de personnes qui ont terminé leurs études secondaires, la population âgée de 15 ans et plus, de 15 à 24 ans et de 55 ans et plus, et le nombre d'actifs dans les domaines suivants: agriculture, exploitation forestière et pêche, industries minières, industries manufacturières, construction, transports et services. L'algorithme utilise aussi des variables relatives au logement comme, par exemple, le nombre total de logements, le nombre de logements loués, le nombre de ménages composés d'une seule personne et le nombre de ménages composés de deux personnes.

#### 4. REMANIEMENT DU PLAN DE SONDAGE DES RÉGIONS AUTOREPRÉSENTATIVES

La taille minimale de l'échantillon des villes classées comme UAR dans le nouveau plan de sondage a été haussée à 50 logements parce qu'une analyse a révélé que les coûts unitaires sont très élevés quand les échantillons des régions AR sont petits. La composition de l'univers des régions AR a toutefois peu changé à cause des effets compensateurs de l'élargissement de l'échantillon de l'EPA de 33,000 à 55,000 ménages à la fin des années 1970 et à cause de la nouvelle répartition de l'échantillon remanié.

Pour les raisons mentionnées plus haut, le plan de sondage des régions AR est resté essentiellement le même, et le principal objectif visé pour ces secteurs était de mettre à jour les mesures de leur taille sans engager les coûts élevés d'un démontrement indépendant sur le terrain comme celui entrepris au dernier remaniement. Les données du recensement de 1981 ont été utilisées pour cette mise à jour, et deux méthodes différentes ont été appliquées aux villes décomposables en côtes d'îlots (grandes villes pour lesquelles on possédait des données agrégées au niveau des côtes d'îlots) et les villes non décomposables en côtes d'îlots. Les techniques de mise à jour et les unités d'échantillonnage choisies pour ces deux méthodes sont décrites plus bas. Le procédé de stratification à deux niveaux adopté pour les régions AR est expliqué à la section 5.

##### Villes décomposables en côtes d'îlots

C'est l'existence de données de recensement agrégées au niveau des côtes d'îlots dans les zones bâties des villes les plus grandes qui a été le facteur clé dans la décision de remanier au complet l'échantillon de ces villes, lesquelles forment les 2/3 de la base des régions AR. Ce remaniement a également permis d'améliorer le schéma de stratification à l'aide du procédé décrit à la section suivante.

Pour les parties de ces villes qui sont décomposables en côtes d'îlots, les îlots de recensement ont été choisis comme grappes (c'est-à-dire comme UPE). On a examiné les composantes de la variance dans un plan de sondage à deux degrés fondé sur la méthode des groupes aléatoires de Rao, Hartley et Cochran pour deux types de grappes-les îlots de recensement et les SD-en simulant le plan de sondage de l'EPA à partir des données du recensement de 1976 pour les UAR d'Hali-fax et de Saskatoon (Choudhry, Drew et Lee 1984). Les résultats de cette étude ont montré que, dans les cas où les mesures de la taille des UAR sont à jour, il y avait peu de différence entre les variances d'échantillonnage obtenues pour les SD et les îlots. Le choix des îlots comme grappes était donc justifié du point de vue opérationnel. Les îlots offraient une base de sondage pres-que toute faite (il a fallu diviser ou combiner des unités dans une proportion de cas variant de 5 à 100%), ce qui permettrait d'élaborer un plan de sondage fortement automatisé et économique à remanier. Autre fait important, les données des recensements futurs peuvent être relevées pour les îlots stables sur le plan géostatistique (mais non pour les SD qui, en tant qu'unités opérationnelles, changent à chaque recensement), ce qui rend possible une mise à jour peu dispendieuse. Les SD ont été choisis comme unités d'échantillonnage dans les parties de ces villes pour lesquelles, changeant tous les cinq ans. Les zones bâties comprennent 86% de ces villes.

Les SD ont été choisis comme unités d'échantillonnage dans les parties de ces villes pour lesquelles on ne disposait pas de données agrégées au niveau des côtes d'îlots. La variance a été calculée seulement pour le cas où les mesures de la taille des régions sont à jour. On croyait que, puisque les SD sont des unités plus grandes que les îlots, elles seraient moins touchées par la croissance très concentrée qui peut avoir lieu à l'extérieur des zones bâties d'une ville. Le choix de SD dans ces régions permettrait également de réduire beaucoup les coûts du remaniement parce que les fractions de sondage sont faibles, peu d'unités ont dû être divisées. L'étude de la variance et l'étude du temps et des coûts (Lemaître 1983) ont révélé que la variance par unité de coût est à peu près constante pour les grappes où on sélectionne de 2 à 8 logements. On a donc décidé de retenir le facteur de densité de 4 ou 5 logements dans chaque grappe, comme dans l'ancien plan de sondage, pour les strates où les grappes sont des îlots, mais de choisir plutôt de 6 à 8 logements dans les SD à cause de leur taille plus grande.

Les grandes UAR étaient stratifiées en profondur par regroupement de secteurs de recensement continus - régions géostatistiques comprenant de 3,000 à 5,000 habitants et dont la stabilité d'un recensement à l'autre en fait des unités opérationnelles pratiques - sans tenir compte de l'optimalité de leurs caractéristiques. Des unités primaires d'échantillonnage, désignées par le terme "grappes" et composées essentiellement d'un îlot urbain, ont été délimitées à partir de comptes relevés sur place en 1973. Un échantillon à deux degrés de grappes et de ménages aléatoires avec probabilité proportionnelle à la taille (ppt). En plus de la base de sondage aérolaire, une base ouverte était tenue à jour pour les immeubles d'appartements dans les grands villages.

La méthode de sélection de la base aérolaire comporte cet avantage particulier que sa souplesse permet de modifier la taille de l'échantillon (Singh et Drew 1977) et en facilite la mise à jour (Platak et Singh 1977, Drew, Choudhry et Gray 1978). L'échantillon des UAR doit être mis à jour périodiquement parce que les comptes utilisés pour la sélection avec ppt deviennent progressivement moins fiables au fil du temps, ce qui accroît la variance d'échantillonnage des estimations de l'enquête. Comme l'échantillonnage se fait de façon indépendante dans chaque groupe aléatoire, l'échantillon peut être mis à jour à l'aide de la méthode de Keyfitz (1957), laquelle permet de réviser les probabilités de sélection à partir de valeurs récentes des comptes de logements, de conserver autant d'unités déjà sélectionnées que possible et d'éviter d'inclure dans l'échantillon mis à jour les logements sélectionnés pour l'ancien échantillon. De 1978 jusqu'au début de la période de remaniement en 1982, des mises à jour régulières ont été faites pour les régions AR en croissance rapide, et presque la moitié de la base de sondage a été modifiée. La fréquence des mises à jour n'a pas été suffisante pour réduire les effets du plan de sondage aux niveaux mesurés au cours des 4 premières années de l'enquête, mais il a été possible d'empêcher de nouvelles détériorations de ces effets qui augmentaient en moyenne de 7% à 8% par année pour les estimations du nombre de chômeurs.

Les unités non autoréprésentatives (UNAR) sont des régions à l'extérieur des UAR et comprennent des régions rurales et de petits centres urbains. À l'intérieur des régions NAR, les régions économiques étaient stratifiées par catégorie d'activité économique de manière à ce que toutes les strates forment des zones de terrain continus. Pour chaque strate, on procédait à une sélection d'unités primaires d'échantillonnage (UPÉ) aussi représentatives de leur strate que possible en ce qui a trait au rapport population rurale/population urbaine et à diverses caractéristiques importantes de la population active. La partie rurale d'une UPÉ était constituée de secteurs de dénombrement (SD) ruraux continus, mais la partie urbaine n'était pas toujours contiguë à cette zone rurale parce qu'il fallait respecter le rapport population rurale/population urbaine. Les plus grands centres urbains dans ces strates étaient souvent inclus dans plusieurs UPÉ à la fois.

Au remaniement de 1973, deux UPÉ ont été sélectionnées dans chaque strate selon la méthode d'échantillonnage systématique avec classement aléatoire et probabilité proportionnelle à la taille (Hartley et Rao 1962). En 1977, la taille de l'échantillon de l'EPA est passée de 33,000 ménages à 55,000 ménages par mois, et cet échantillon agrandi a été réparti de manière à améliorer la fiabilité des données provinciales. Ainsi, la taille de l'échantillon affichait une hausse proportionnelle plus marquée dans les petites provinces. L'échantillon des régions NAR a été élargi par l'addition de 1 à 4 UPÉ supplémentaires à chaque strate (Gray 1975).

Les parties rurales et urbaines des UPÉ sélectionnées étaient échantillonnées indépendamment. Dans les parties urbaines de ces UPÉ, un échantillon à deux degrés composé de grappes et de logements était prélevé, tandis qu'un échantillon à trois degrés composé d'unités secondaires d'échantillonnage (SD du recensement ou combinaisons de SD), de grappes (zones de terrain faciles à délimiter comptant au plus 20 logements) et de logements. Sauf pour la dernière étape de l'échantillonnage, la sélection des unités était fondée sur la méthode d'échantillonnage systématique avec classement aléatoire et ppt.



## 2. DÉFINITION DES OBJECTIFS

L'une des étapes fondamentales du remaniement d'une enquête régulière est la redéfinition des objectifs de l'enquête. Pour l'EPA, il a fallu réévaluer non seulement son rôle de principale source de renseignements à jour sur le marché du travail, mais également son utilisation à l'intérieur de Statistique Canada comme instrument général pour d'autres enquêtes sur les ménages (Singh et Drew 1981b).

Vers le début du programme de remaniement, il a été décidé que celui-ci devrait avant tout viser l'amélioration de l'EPA, mais qu'il était également important d'accroître la flexibilité de cette enquête pour des applications générales. À cet égard, plusieurs changements sont en cours et auront des effets positifs non seulement sur l'EPA, mais aussi sur les autres enquêtes qui y sont liées. Les objectifs concernant la fiabilité des données sur le marché du travail sont décrits plus bas et les modifications qui amélioreront les autres enquêtes sont résumées à la section 9.

Les objectifs de production de données sur le marché du travail ont été établis, lors de conférence avec les points de contact statistiques dans les dix provinces du Canada et avec les principaux ministères fédéraux qui utilisent ces données. En général, ces consultations ont permis de constater une satisfaction à l'égard de la fiabilité des données provinciales et nationales, mais un grand besoin de données infraprovinciales plus précises. Les objectifs suivants ont été fixés pour la production de données fiables à partir de l'échantillon remanié:

- (i) pour le Canada et chacune des dix provinces, éviter toute réduction de la fiabilité des estimations globales des niveaux et des variations mensuelles de l'emploi et du chômage;
- (ii) pour les 24 régions métropolitaines de recensement définies pour le recensement de 1981, produire des estimations du nombre de chômeurs dans les régions où le coefficient de variation (CV) est inférieur ou égal à 20%;
- (iii) pour les 66 régions économiques infraprovinciales définies en consultation avec les provinces, produire des estimations mensuelles du nombre de chômeurs dans les régions où le CV est inférieur ou égal à 25%;
- (iv) pour les villes de 60,000 habitants ou plus au Québec et en Ontario et celles de 25,000 habitants ou plus dans les autres provinces, produire des estimations trimestrielles du nombre de chômeurs dans les villes où le CV est inférieur ou égal à 25%;

Pour atteindre ces objectifs, il a fallu redistribuer une partie de l'échantillon des grandes villes et régions économiques dans certaines villes et régions économiques plus petites. Comme on voulait en plus réduire le coût de l'EPA (section 8), beaucoup d'espaces reposaient sur les projets de recherche concernant les moyens d'accroître l'efficacité des opérations de collecte des données et de production des statistiques de l'EPA. Dans les sections suivantes, ces problèmes sont examinés du point de vue des régions autorensementales (AR) et des régions non autorensementales (NAR).

## 3. DESCRIPTION DE L'ANCIEN PLAN DE SONDAGE DE L'EPA

Une description complète de l'ancien plan de sondage de l'EPA a été présentée par Platak et Singh (1976). Cette section résume les principaux aspects de ce plan de sondage afin de rendre plus clairs les sujets traités dans les sections suivantes.

Les unités autorensementales (UAR) de l'ancien plan de sondage correspondaient aux villes qui, au moment de la conception du plan de l'enquête, étaient assez grandes pour avoir comme rendement prévu un échantillon de 20 ménages, soit le nombre minimal de ménages pouvant être confié à un interviewer. La limite inférieure de la taille des UAR variait de 10,000 personnes dans la région de l'Atlantique à 25,000 personnes au Québec et en Ontario.

# Remaniement de l'enquête sur la population active au Canada à partir des résultats du recensement de 1981<sup>1</sup>

M.P. SINGH, J.D. DREW et G.H. CHOUDHRY<sup>2</sup>

## RÉSUMÉ

Après chaque recensement décennal de la population, l'échantillon de l'enquête sur la population active au Canada (EPAC) est remanié pour tenir compte de l'évolution des caractéristiques de la population et répondre aux nouveaux besoins en information. Le dernier programme de remaniement, qui a amené la sélection d'un nouvel échantillon au début de 1985, comportait des recherches poussées sur les moyens d'améliorer le plan de sondage, la collecte des données et les méthodes d'estimation. Les grandes lignes de ce programme sont décrites ici.

MOTS CLÉS: Enquête permanente; plan de sondage à plusieurs degrés; stratification; redistribution d'un échantillon; interviews téléphoniques; méthode itérative du quotient.

## 1. INTRODUCTION

L'enquête sur la population active au Canada (EPA), la plus vaste enquête mensuelle sur les ménages menée par Statistique Canada, est ordinairement remaniée après chaque recensement décennal. Dans le cadre du remaniement entrainé après le recensement de 1981, un programme intensif de recherche que nous avons décrit dans une autre étude (Singh et Drew 1981a) a été mis en oeuvre pour examiner diverses méthodes d'échantillonnage, d'estimation et de collecte des données. Comme les données sur le marché du travail sont assez fiables au niveau national et provincial, les principaux objectifs fixés pour ce remaniement étaient d'améliorer la fiabilité des données infraprovinciales et d'accroître le rendement général de l'enquête. Pour augmenter le rendement, l'accent a été mis sur les moyens d'automatiser davantage les diverses phases de l'échantillonnage, une utilisation accrue des données du recensement pour mettre à jour l'échantillon au lieu d'informations recueillies indépendamment et un recours accru aux interviews téléphoniques dans le déroulement normal de l'enquête. Pour améliorer les données infraprovinciales, diverses techniques d'échantillonnage et d'estimation ont été examinées, et les résultats ont entrainé des modifications dans les méthodes utilisées auparavant et une redistribution de l'échantillon à l'intérieur des provinces.

Cet exposé résume les résultats des recherches théoriques et empiriques et des expériences sur le terrain entreprises au cours du programme de remaniement. Les sections 2 et 3 décrivent les objectifs du remaniement et l'ancien plan de sondage, après quoi les sections 4, 5 et 6 présentent les conclusions tirées des recherches qui ont justifié les modifications apportées aux méthodes d'échantillonnage et de collecte des données. La section 7 porte sur les problèmes d'estimation et la section 8 sur la redistribution de l'échantillon. Les principales répercussions du remaniement de l'échantillon sur d'autres enquêtes liées à l'EPA sont soulignées à la section 9 et, enfin, la section 10 aborde brièvement les avantages découlant des principales améliorations de l'échantillon remanié et mentionne quelques futurs projets de recherche.

<sup>1</sup> Présenté à une réunion de la Section sur les méthodes d'enquête de l'American Statistical Association à Philadelphie, août 1984.

<sup>2</sup> M.P. Singh, J.D. Drew et G.H. Choudhry, Division des méthodes de recensement et d'enquêtes-ménages, Direction de la méthodologie, Statistique Canada, 4 étages, Immeuble Jean-Jalon, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6.





en reconnaissance de sa contribution en tant que membre fondateur  
du comité de rédaction jusqu'à sa retraite.

**À M. Paul Francis Timmons**

# TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

## COMITÉ DE RÉDACTION

Président R. Platek Statistique Canada

Rédacteur en chef M.P. Singh Statistique Canada

Rédacteurs associés K.G. Basavarajappa

D.R. Bellhouse Statistique Canada

E.B. Dagum Statistique Canada

J.F. Gentleman Statistique Canada

G.J.C. Hole Statistique Canada

T.M. Jeays Statistique Canada

G. Kalton Statistique Canada

C. Patrick Statistique Canada

J.N.K. Rao Statistique Canada

C.E. Särndal Statistique Canada

V. Tremblay Statistique Canada

H. Lee Statistique Canada

R. Platek (Président), E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

## POLITIQUE DE RÉDACTION

La revue *Techniques d'enquête* publie des articles sur les divers aspects des méthodes statistiques- qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les techniques de lissage et d'extrapolation, les études démographiques, l'intégration et l'analyse de production de statistiques. Une importance particulière est accordée à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles sont soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue *Techniques d'enquête* est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, 4<sup>e</sup> étage, Edifice Jean-Talon, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

# TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada  
Volume 10, numéro 2, décembre 1984

## TABLE DES MATIÈRES

M.P. SINGH, J.D. DREW, et C.H. CHOUDHRY	Remaniement de l'enquête sur la population active au Canada à partir des résultats du recensement de 1981	139
D.A. BINDER, M. GRATTON, M.A. HIDIROGLOU, S. KUMAR, et J.N.K. RAO	Analyse de données qualitatives d'enquêtes complexes: quelques expériences canadiennes	155
P.A. CHOLETTE	Estimateurs des cycles économiques dans les séries semestrielles	171
J. COULTER	Appariement d'enregistrements pour l'évaluation des erreurs non dues à l'échantillonnage dans le recensement de l'agriculture de 1981 au Canada	179
A. CHAUDHURI et R. MUKERJEE	Estimation sans biais des paramètres d'un domaine dans l'échantillonnage sans remise	197
D. DOLSON, P. GILES, et J.-P. MORIN	Méthode d'enquête sur les personnes souffrant d'une incapacité à l'aide de questions supplémentaires de l'enquête sur la population active	203
	Rectification	215
	Remerciements	217





# TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA

Décembre 1984

Publication autorisée par  
le ministre des Approvisionnements et  
Services Canada

© Ministre des Approvisionnements  
et Services Canada 1985

Avril 1985  
8-3200-501

Prix: Canada, \$10.00, \$20.00 par année  
Autres pays, \$11.50, \$23.00 par année

Paiement en dollars canadiens ou l'équivalent

Catalogue 12-001, vol. 10 n° 2

ISSN 0714-0045

Ottawa

Bon de commande

Poster à l'adresse: Vente et distribution des publications

Statistique Canada  
Ottawa (Ontario)  
Canada KIA 0T6

Veuillez accusé réception de mon abonnement à la revue Techniques d'enquête (cat. no. 12-001). Le prix est de 10,00\$ par copie, 20,00\$ par année au Canada, et de 11,50\$ par copie, 23,00\$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent).

Je suis présentement abonné(e) à une ou plusieurs publications de Statistique Canada:

☐ Oui. Mon numéro d'abonné(e) est \_\_\_\_\_ (Veuillez attendre la facture avant de payer)

☐ Non. Votre commande ne peut être traitée qu'en complétant un formulaire de commande ou en cochant une des cases de mode de paiement ci-dessous.

Numéro de commande \_\_\_\_\_ (S.V.P. inclure votre numéro de commande.)

Nombre de copies \_\_\_\_\_ Montant envoyé \$ \_\_\_\_\_

☐ J'inclus mon paiement fait à l'ordre du Receveur général du Canada/publications. (N° INTRA 0540) (Compte Crediteur n° 0051)

☐ Veuillez porter à mon compte \_\_\_\_\_ de Statistique Canada.  
☐ Veuillez porter à mon compte ☐ VISA ☐ MASTERCARD.

Numero de la carte \_\_\_\_\_

Date d'expiration \_\_\_\_\_

Nom du titulaire \_\_\_\_\_ (en lettres moulées)

Banque émettrice \_\_\_\_\_

Signature du titulaire \_\_\_\_\_

Veuillez expédier à (en lettres moulées):

Nom (Organisme) \_\_\_\_\_

Division \_\_\_\_\_

A l'attention de \_\_\_\_\_

Adresse \_\_\_\_\_

Ville \_\_\_\_\_

Province \_\_\_\_\_

Code postal \_\_\_\_\_

Téléphone \_\_\_\_\_





# TECHNIQUES D'ENQUÊTE

UNE REVUE  
DE  
STATISTIQUE CANADA

VOLUME 10, NUMÉRO 2  
DÉCEMBRE 1984

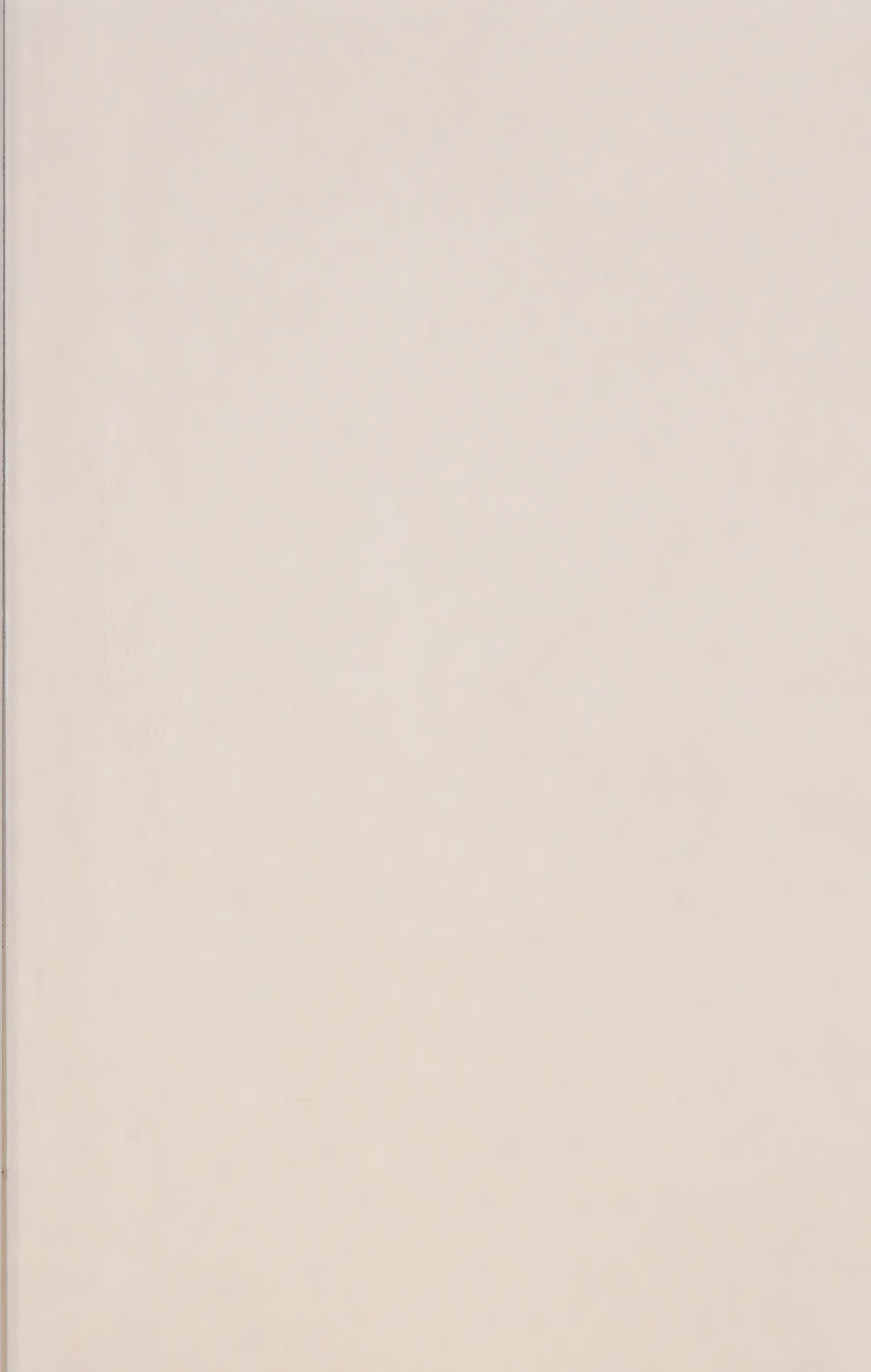
Canada













JUN 10 1987



